DOCUMENT RESUME

ED 151 615 CE 015 679

TITLE Emergency Medical Services: Research Methodology. Research Proceedings Series.
INSTITUTION National Center for Health Services Research and Development (DHEW/PHS), Rockville, Md.
PUB DATE Dec 77
NOTE 133p.; Proceedings of a conference (Atlanta, Georgia, September 8-10, 1976)

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS Administration; Administrative Problems; Administrator Role; Conference Reports; Economic Research; Emergency Programs; Field Check; *Medical Care Evaluation; *Medical Services; Program Effectiveness; *Program Evaluation; Research Design; *Research Methodology; Social Science Research
IDENTIFIERS *Emergency Medical Service

ABSTRACT

The thirteen papers included here were presented at a conference on the importance of systematic research in evaluating the Emergency Medical Services (EMS) system and administrative functions. The first paper spells out the roles and responsibilities EMS administrators incur when they make a commitment to participate in a research project. An additional paper attempts to explain the role that research data can play in the resolution of operational problems in service agencies. Another paper describes the types and levels of evaluation research being done. Papers are also presented on experimental design, attitude measurement, the development of indicators of program effectiveness, measuring the monetary value of lifesaving programs, randomized field tests, and putting together evaluation research staff. Finally, several papers which deal with research conducted in a police setting are included to offer an instructive analogy to the conflicting yet mutually dependent roles of administrator and evaluator which are also present in EMS services. (BB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151615

NCHSR

RESEARCH PROCEEDINGS SERIES

Emergency Medical Services: Research Methodology

Proceedings of a conference held in Atlanta, Georgia, September 8-10, 1976

December 1977

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
National Center for Health Services Research

DHEW Publication No. (PHS) 78-3195

ED 015679

National Center for Health Services Research Research Proceedings Series

The *Research Proceedings Series* is published by the National Center for Health Services Research (NCHSR) to extend the availability of new research announced at conferences, symposia and seminars sponsored or supported by NCHSR. In addition to publishing the papers given at key meetings, this series includes discussions and responses whenever possible. The series is intended to help meet the information needs of health services providers and others who require direct access to concepts and ideas evolving from the exchange of research results.

Abstract

The focus of this conference, held September 8–10, 1976, in Atlanta, Georgia, was the importance of systematic research in evaluating the Emergency Medical Services system, and administrative functions. Presentations made at the conference and compiled in this document deal with a range of conceptual and methodologic issues. Particular attention is given to the opposing yet mutually dependent roles of the administrator/evaulator. Several papers presenting aspects of research conducted in a police setting offer an instructive analogy to emergency medical services systems.

Foreword

The Emergency Medical Services Systems Act of 1973 (P.L. 93–154) established comprehensive regional emergency medical services (EMS) systems in an attempt to integrate a number of public and private services, including communications, transportation, personnel and facilities, into coordinated programs designed to save lives and to reduce disability. In the 1970s, however, we are moving from pursuing health goals "at any price" to a realization that our resources are limited, and we must make deliberate choices. The goal of "the best for everyone" provides no guidance for deciding among alternative system designs and alternative uses of scarce resources. The EMS Systems Act focuses on improving the effectiveness of emergency services; a growing national concern for containing the rapidly-rising costs of health care introduces the requirement that system efficiency be considered as well.

This conference, held in Atlanta, Georgia, September 8–10, 1976, assessed the value of research methods in analyzing and evaluating EMS systems. The conference emphasized the critical role of the system administrator as both a facilitator and a user of evaluative research. In addition to conceptual and methodologic presentations, a group of papers presented an analog case study of the collaboration between The Police Foundation and the Kansas City Police Department. Recurring throughout are references to the conflicting, yet mutually dependent, roles of administrator and evaluator. The police analogy offers an example of successful, if precarious, resolution of those two roles and of the insider-outsider viewpoints. Police work is not emergency medical services work, but the questions of "what difference does it make?" and "what makes a difference?" are there for both public services, and there are operational and political considerations, technology, and evaluative measures (e.g., response time) which are common to both systems. The problems and motivations of the police administrator may offer new insights and approaches for the EMS administrator.

Gerald Rosenthal, Ph.D.
Director
December 1977

Contents

iii FOREWORD v
Gerald Rosenthal
National Center for Health Services Research

1 INTRODUCTION
Lee Sechrest
Florida State University

3 Administrative Functions and Research Requirements
Lee Sechrest
Florida State University

6 Research in The Context of Delivery of a Critical Public Service: The Kansas City, Missouri Police Department Experience
Major Lester N. Harris
Kansas City, Missouri Police Department

16 Evaluation Results and Decision-Making: The Need for Program Evaluation
Lee Sechrest
Florida State University

24 Evaluation Research: What Is It and How Is It Done?
Linda Victor Esrov
Florida State University

33 Experimental Design and Causal Inference
Lee Sechrest
Florida State University

45 Social Attitudes and Program Evaluation
Russell D. Clark, III
Florida State University

53 Recruitment, Selection, Training and Supervision of Civilian Observers to Work in Police Patrol Operations Research
William Bieck
Kansas City, Missouri Police Department

62 Developing Indicators of Program Effectiveness: A Process
George L. Kelling
Police Foundation

67 Measuring The Monetary Value of Lifesaving Programs
Jan Paul Acton
The Rand Corporation

84 Economic Analysis and the Evaluation of Medical Program
Jan Paul Acton
The Rand Corporation

vi

87 Appropriateness and Feasibility of Randomized Field Tests
Robert F. Boruch
Northwestern University

105 Development of Staff for Evaluations (A RETROSPECTIVE VIEW)
George L. Kelling
The Police Foundation

115 Evaluation of Experiments in Policing: What Are We Learning?
Joseph H. Lewis
The Police Foundation

124 Biographical Sketches

Introduction

It has become widely apparent that at least a part, and often a large part, of poorly planned and implemented program evaluation research is the inhospitable climate that exists for such research in many systems and organizations. The climatic insufficiency may involve lack of understanding of the need for evaluation, outright hostility to evaluation, or a lack of appreciation for the conditions required for a good evaluation to take place. If program administrators do not look favorably on evaluation, it is virtually certain that even if evaluation is attempted, it will be unsuccessful. However, even enthusiastic program administrators may obstruct, impede, and destroy evaluation attempts for want of understanding of the rather demanding conditions which must be met in order for evaluation research to succeed. Numerous other writers have made the same and additional points on the topic (e.g., Campbell, 1969, 1975a, 1975b; Gurel, 1975; Rivlin, 1971; Weiss, 1970, 1972, 1973, 1975).

A part of the problem that administrators have with evaulation research undoubtedly stems from their perceived vulnerability to potentially unfavorable outcomes, vulnerability that is often enhanced by their very own promises about what a program will produce, by what Campbell (1969) calls the *overadvocacy trap*. However, not only may administrators be less vulnerable than they suppose, with a really good understanding of the nature and purposes of evaluation, they might come to see it as a potentially valuable tool to be used in the accomplishment of successful programs. With a better understanding of why and how good evaluation research is carried out, administrators might also be less likely to impede or subvert the research by decisions made in relation to it. For example, they might be more willing to plan for strong evaluation in the first place, to provide the resources necessary to carry out the evaluation, to refrain from operational changes that would drastically affect the evaluation, etc. The view here is that program administrators are not necessarily and inherently the enemies of evaluators, with their informal cooperation good evaluation research is difficult to achieve; without it, good evaluation research is impossible.

Another point which might be made by way of background is that research, such as it is, into the factors affecting the utilization of research by policy makers points to the importance of involving policy makers in the research whose results are to be applied (e.g., Havelock, 1969; Salasin & Davis, 1975). Not only does involvement of administrators in ongoing research result in a degree of co-opting that might make them more interested in the findings, but they also may have a greater appreciation of the nature and potential use of the results by having had a hand in producing them.

Clearly there is a need for high quality research in emergency medical services. Yet there is a dearth of proposals of any quality at all. While the reasons for lack of good EMS proposals are undoubtedly complex, perhaps in some degree being inherent in the nature of the problems, there is no question that a good part of the problem stems from lack of research talent in EMS systems. There may be additional problems resulting from a lack of strong commitment to doing research in the first place. Because of the importance of emergency services in the overall system of health care in this country, EMS would seem to be an appropriate area in which to attempt a general upgrading of research efforts, including the planning and preparation of proposals.

While there are several possible levels at which one might try to intervene in EMS in order to improve research, e.g., research workers already in the field, Regional EMS offices, etc., the conference was directed toward persons currently involved in the operations of emergency medical services at some level. The aim was to attract administrators with operational and decision-making responsibilities on the grounds that these persons are in a position to facilitate good research if they understand the need for it and the requirements of research that may infringe upon administrative functions.

Although many of the requirements for good quality research may be formulated in the abstract,

i.e., without reference to particular fields or content, emergency medical services seemed to be a sufficiently complex potential research area to justify a conference focusing specifically on it. However, there has as yet been relatively little research at all on emergency medical systems and even less that may be presented as exemplary. On the other hand, there has been in the past few years a rather surprising quantity of good quality research on police practices. There are many similarities in the research problems likely to be encountered in police and emergency medical services research since both involve the delivery of a critical public service, often under considerable pressure. Both of them are *public* services, i.e., they cannot choose their clientele, and both of them involve delivery of services by individuals with less than professional education and training, and typically by persons with no more than high school education.

In view of the above considerations it seemed potentially worthwhile to involve in the EMS conference a number of persons with experience in the police research field. There was no thought that any kind of simple correspondence could be made between police and EMS systems, but it was thought that the experiences in police research would be relevant and instructive. Since the Kansas City, Missouri, Police Department has been involved in some of the largest and most innovative police research projects, participation of individuals associated with those projects was solicited. In addition, it was believed that the experiences of The Police Foundation, which has funded and monitored much of the police research work which has been done recently, would be of great interest.

The aim of the conference was not to make researchers out of administrators, but to try to convey a sense of the importance of systematic research and of the nature of research, especially as it relates to operational and administrative functions and goals. The topics chosen for the papers were meant to reflect a range of views and issues, hopefully in a way quite meaningful and comprehensible to EMS administrators. The papers were not intended for use by the professional research community.

The panelists who made presentations at the conference were:

Lee Sechrest, Ph.D., Research methodologist, Florida State University, Conference Director.
Robert Boruch, Ph.D., Research methodologist, Northwestern University.
Jan Acton, Ph.D., Economist, RAND Corporation.
William Biech, Project Director, Response Time Analysis Study, Kansas City, Missouri, Police Department.
Russell D. Clark, III, Ph.D., Social psychologist, Florida State University.
Linda Esrov, Ph.D., Research methodologist, Florida State University.
Lester Harris, Major, Kansas City, Missouri, Police Department.
George Kelling, Ph.D., Sociologist, The Police Foundation.
Joseph Lewis, Executive Director, The Police Foundation.
Robert Thorner, D.Sc., National Center for Health Services Research.

The Conference agenda was approximately as follows:

Introductory remarks.
Priorities in emergency medical systems research.
Evaluation results and decision making: the need for program evaluation.
Types and levels of program evaluation.
Problems in causal inference.
Evaluation experiment simulation exercise.
Research in the context of delivery of a critical public service.
Measuring the outcomes of social programs.
Direct and indirect outcome measures.
Program assessment simulation exercise.
Social attitudes and program evaluation.
Cost benefit and cost effectiveness.
Simulation exercise and discussion.
Project administration and data quality control.
Examples of good evaluations.
The politics of evaluation and implementation of findings.
Putting together a good evaluation research team.
Funding of research on emergency medical systems.

Administrative functions and research requirements

Lee Sechrest
Professor of Psychology
Florida State University
Tallahassee, Florida

The first paper spells out the roles and responsibilities administrators incur when they make a commitment to participate in a research project. A similar and equally demanding paper could be written about the responsibilities of researchers working in a service delivery setting. It should not be thought that the problems and the shortcomings are all on one side.

In order for good quality research to be planned and carried out, it is essential to have full support from administrators in agencies involved in the research. That statement might seem so obvious as not to need utterance, but it is unfortunately the case that, however obvious the principle, all too frequently the quality of research suffers drastically because of lack of administrative commitment and support. In part the problems may stem from failure of administrators to understand research needs, in part from a failure to understand what they are really getting into in beginning a research project, in part from the inexorable political and public pressures that surround the delivery of all critical service, and in part the problems clearly stem from failure of researchers to understand the service delivery context and the administrative role. We want here to clarify as much as possible the way in which administrative functions impinge upon research. It is our expectation that in at least some degree to be forwarned is to be forearmed, and perhaps with better understanding on both sides of what is involved in doing quality research in a service delivery setting, at least some of the difficulties and perhaps most of the disasters can be obviated.

There is, to begin with, the recognition that a problem exists and that systematic research might provide information useful in solving the problem. It has been evident to many of us involved in action research settings that problems are not always equally well recognized by administrators and researchers, are not necessarily defined in the same terms by both groups, and that a conviction of the probable value of systematic research is often lacking in administrators. Researchers tend for obvious reasons to have a broader perspective on problems than do most administrators. Researchers tend to be concerned with the general case while administrators are concerned with their own particular agency. Consequently in some instances a researcher may see and want to work on a problem which simply does not exist in or is not of concern in the setting in which the work is to be carried out. Researchers may, for example, be interested generally in the relationship between training and performance of medical personnel while in a particular health delivery setting that concern may be minimal, perhaps even justifiably minimal. An administrator may see a problem as involving limited resources with which to work while a researcher might prefer to define the problem in terms of optimizing distribution of resources available. When there is not a congruent recognition and definition of the problem to be worked on by both administrators and researchers, trouble is at hand. There will be a differential commitment to the research, different notions about its goals and how to reach them, and discrepant views of the importance of research as opposed to administrative priorities in subsequent decision making.

Clearly, then, a first step in the planning of the research process is to ensure that administrators and researchers have a common definition of the problem. If the research is directed toward a problem of general interest, perhaps one involving fundamental principles rather than immediate and local concerns, it is important that the administrator recognize as well as the researcher the need for work on the problem and the perhaps somewhat altruistic contribution that his agency will be making. Without that equality of recognition and commitment to the idea of research as a sort of societal obligation, researchers and administrators are bound to clash when, as is inevitable, the pressures of operational problems begin to lead to changes in procedures that will weaken or even ruin the research. And administrator who is "talked into" participating in a project in which he has no interest or for which he feels no enthusiasm is making a mistake in ever beginning the project.

When a research project is planned and undertaken in an action setting, there are a great

many implicit restrictions on the freedom of an administrator to operate in a normal manner. It is highly desirable that these restrictions be made explicit and that they be discussed frankly and fully between the researcher and the administrator. It may even be a good idea to write them down and have both parties initial the document to which they are agreeing. Unfortunately, it is not always the case that the restrictions are recognized in advance by either party, and if they were recognized, probably a good many projects would not be undertaken—which might be for the best.

The specific restrictions that may be implicit in research plans will differ somewhat from one project or one setting to another, but there are some common ones that can be stated. First, if the project is an experiment, certainly if it is a true experiment and often even if it is only a quasi-experiment, an administrator will be restricted by the plan in the terms of which services can be delivered. The design of the experiment may call for some persons to receive services while they are withheld from others, or for different persons to receive different services, and the administrator may not be allowed to participate in that decision. Treatments may be allocated randomly to people (or to whatever units are involved), they may be allocated serially, or in any one of many other possible ways. It is a potentially severe restriction on an administrator's authority if he or she cannot decide to whom or how services are to be delivered.

A particularly troublesome problem for most administrators arises in those cases in which the research design calls for withholding of some form of treatment for some cases. Even though the very reason for doing the research in the first place may be that the effectiveness of a treatment is open to serious question, it may still be difficult for an administrator in a politically sensitive setting to make the decision—and then to stick by it—of letting some cases go without treatment or be exposed to what is feared to be an inferior treatment. As is pointed out elsewhere in these papers, there is a powerful tendency for the effectiveness of treatments to become *assumed* before there is any evidence. Nonetheless, on occasion the administrator wishing to pursue a certain line of research may have to steel himself to the risk of a "no treatment" control group. Once the commitment is made, it is important that it be adhered to until the evidence is firm one way or another. The costs of mounting an experiment at all are usually too high to think of having them aborted.

It should also be understood that there are similar restrictions on other persons actually delivering the services, and one of the tasks of the administrator may be to assist in enforcing the experimental plan. Physicians may have to be told, for example, that treatment plans are to be followed even when it goes against their own personal inclinations or even judgment. In a study of the value of diverting certain juvenile offenders from the criminal system it was found that some police officers were using their knowledge of the serial process by which juveniles were being assigned either to diversion or to custody to gain the type of treatment they thought best for particular kids they worked with. The police officers were getting into the records files after hours and changing the order of the cases that would be assigned the next day. In becoming involved in a research project, administrators assumed at least some responsibility for the scientific integrity of the project. Even under the best of circumstances it is difficult to maintain randomization of treatments, and an administrator can be of great help if he determines that the experimental plan will be carried out.

Administrators also very often lose freedom with respect to at least some of the characteristics of the treatment that is being administered. In particular the freedom to make changes or other adjustments in the form of the treatment may be sacrificed for the duration of the experiment. It is readily apparent to most people why in the course of testing a drug it is impermissible to change the drug in any way during the trial. It is seemingly more difficult to see and accept, but it is equally impermissible to change other treatments during the time they are being tested. If one wanted to test the efficiency of some type of emergency room organization against an alternative, one would need to decide from the beginning what the new organization should be and then stick with it fairly closely until results became conclusive. One could not, without seriously jeopardizing the interpretation of the experiment, continue to organize and reorganize. Again, the point may seem obvious, but it becomes a sticky issue repeatedly when research is being done in action settings. In the Kansas City police patrol experiment (Kelling et al., 1974), about which more will be said later, it was regarded as of utmost importance that different patterns of patrol be effectively maintained in the experimental and control areas. However, because one pattern being tested went so much against the grain of current police beliefs and practices, there were constant threats of subverting the treatment plan, e.g., by patrolmen entering areas on their own initiative. It required utmost attention from both the experimenters *and* police administrative officials to maintain the conditions of the experiment reasonably well. The integrity of the experimental treatment is also at least a partial responsibility of the administrator.

Operational procedures not directly part of the experiment may also need to be kept constant during the course of the study. Record keeping systems, for example, should not be changed

midstream. In the Hawaii Experimental Medical Care Review Organization (1973), as an especially informative example, a system was established to do peer review of treatment of target diseases in hospitals. Data were available for a baseline period and then for the period following the beginning of peer review. Unfortunately, at the time peer review started there was also a critical change in the wording of one requirement, making it more likely that it would have been met and therefore that peer review would appear effective. Administrators making the decision to participate in a research project may also find themselves being called upon to change their record keeping or data collecting procedures and then find that they are in some degree responsible for data quality control. The requirements of the research may necessitate the keeping of records that would not ordinarily be kept, and the maintenance of data quality control may involve extra monitoring of various persons and processes. These matters should be well understood and worked out before the research begins. Geographical boundaries for service districts may have to be kept the same even though strictly operational considerations would dictate a change. Even changes in personnel may have to be avoided if an experiment is going to produce convincing results. It is worth remembering that an unpersuasive scientific investigation is a waste, and the appearance, as well as the actuality, of objectivity and integrity is important.

Finally, of course, administrators may experience a subjective sense of loss of budgetary control within their organizational units. The budget allocated for research may sometimes seem quite large in relation to the operational budget, and the administrator can find her or himself in a situation in which a lot of money is being spent by a lot of people in ways that are threatening. That threat will be especially likely if some of the administrator's own staff become part of the research project or that the basis for their professional loyalties seems distinctly shifted. There are a lot of research projects in which an administrator is likely to come to wonder just what is in it for him or her.

Again we can offer no panaceas. In the best of situations the administrator and the researchers will have a sense of colleagueship, of being embarked *together* on an important and ultimately rewarding venture. That sense of joint responsibility and cooperativeness is best fostered by an open relationship from the beginning, one in which each participant has a good understanding of the other's problems and intentions and in which each has a firm commitment to the same goals.

One factor limiting the participation of an administrator in a research project may well be the doubt of the administrator that anything of value is likely to be gained by the research. For one thing, very few administrators of any kind are trained in research, so that they do not understand it and have little appreciation of how it is done and what it may have to offer. There is, in fact, a type of administrative style widely taught and admired in which an administrator engages in a period of "fact finding," depending largely upon subordinates for input, and then enters his inner-office for a period of mulling things over before announcing a personal, and correct, decision. Preferably the period of mulling over should be brief so as to maintain a reputation for decisiveness. Research which has been done to date on the utilization of scientific and other information in decision making processes indicates that far too many administrators and managers are interested in research findings only if they confirm what is already believed. Administrators also tend to have limited trust in any research that was not done within their own organizations. That mistrust not only narrows drastically the information sources which are searched, but it may also make many administrators doubt the value of doing research of a more basic or general nature that does not bear directly on a problem of immediate interest.

The foregoing warnings and stringent prescriptions for working out everything in advance should not be taken as indicative of the near impossibility of doing good research at all in an action setting. Rather they are meant to be realistic assessments which, if taken into account, can make the difference between a good research project and a failure. Not all of the problems referred to are likely to be relevant in any one setting, and some of the others can probably be rather easily resolved. Nevertheless, the problems of doing good quality research should not be underestimated. As will be evident from examples presented throughout these papers the problems are formidable and not likely to be solved satisfactorily by any team who approach the research task with the idea that it is going to be easy. It almost never is. But it is not so difficult as to be effectively impossible.

References

Kelling, G.L., Pate, T., Dieckman, D., & Brown, C.E. *The Kansas City Preventive Patrol Experiment: A Technical Report.* Washington, D.C.: The Police Foundation, 1974.

The Hawaii EMCRO: An experiment in Non-punitive peer Review. Project Report. Grant No. 5 R18 HS–00795 SRC. National Center for Health Services Research, Rockville, Maryland, 1973.

Research in the context of delivery of a critical public service: the Kansas City, Missouri Police Department experience

Major Lester N. Harris
Kansas City, Missouri Police Department

6

Major Lester Harris has spent all his career as a policeman in the Kansas City, Missouri Police Department. He had overall responsibility for the Kansas City police patrol experiment and has been heavily involved in the other research projects carried out and underway in that Department. Major Harris was asked to discuss the ways in which research came to be a regular activity of his Department and the problems that are involved in carrying out research while at the same time having responsibility for public services of a highly visible and even critical nature. It is believed that there are enough similarities between the politics, structures, and missions of a major police department and an emergency medical system or rescue service to make the lessons from the police department instructive.

Prior to beginning a description and discussion of research and planning within an organization, at least a general description of that organization is needed in order to provide some context for the information.

The Kansas City, Missouri Police Department, unlike the vast majority of municipal police departments, is not under the administrative control of the city government. The department is under state control, operating and administered under the provisions of Missouri Statutes, sections 84.350 through 84.890. Under provisions of these statutes, the governor of the State of Missouri, with consent of the senate, appoints four citizens of Kansas City as a Board of Police Commissioners. The Mayor of the City, by virtue of his office, is the fifth member of the Board. The power and responsibility of police service is vested in this Board of Police Commissioners. The Chief of Police is appointed by the Board of Police Commissioners and serves at the pleasure of the Board. The statutes define the powers and responsibilities of the Board and the Chief, set forth rank structure and salary ranges, addresses personnel administration matters, defines arrest powers, sets forth budgeting and fiscal provisions, etc.

Within the police department, the topmost or largest organizational entities are termed "bureaus." Presently there are four bureaus; Operations Bureau, Administration Bureau, Investigations Bureau, and Services Bureau. The bureau commanders report to the Chief of Police and, with the exception of a few functions, whose heads report directly to the Chief of Police (e.g., Media Liaison and Legal Advisor), all organizational elements are a part of and subordinate to one of the bureaus. The next level of organizational elements below bureaus are called "divisions" and they are in turn divided into "units." The organizational structure is not considered to be sacred or permanently fixed; it is only a framework within which resources are organized in order to facilitate coordinated efforts toward departmental objectives. Alterations are made in the organizational structure as needed.

Department personnel strength is presently 1,212 law enforcement personnel, 479 full time regular civilian personnel, 85 part time school crossing guards (during the school year), 48 temporary contract civilian personnel, and 102 reserve police officers.

Kansas City, Missouri is a city of 316.83 square miles with a 1970 census population of 507,409. It is the principal municipality of a metropolitan area with a 1970 census population of 1.4 million. There are parts of three counties within the city limits, and the western city limit is comprised of the Missouri-Kansas state line. In 1975 there were 46,530 Part I criminal offenses reported to the police department.

The department first established a Planning and Research Unit in about 1953. The unit at that time, and for the following decade, was staffed with only two or three officers. This unit maintained a small departmental library and compiled necessary information and statistics for a department annual report, which is required by state statutes. The resources and efforts of the unit were otherwise involved essentially in routine staff studies and the development and writing of procedures as directed by the Chief of Police or necessitated by current demands on the department. For example, during the period of 1957 through 1963, the city of Kansas City, Missouri annexed a total of 235 square miles in annual increments. The Planning and Research Unit performed much

of the staff work necessary to expand police service to these annexed areas each year, such as determining patrol beat boundaries, etc. The unit did very little work which could properly be termed as experimental or innovative in nature. Our department was certainly not alone in this regard, however. The prevailing attitude in policing was that "if it has worked for the past twenty years there's no reason to change it."

One notable exception to this relatively non-progressive stance was the introduction in 1953 of patrol cars manned by only one officer instead of the traditional two officers. This very significant departure from tradition was implemented in the Kansas City, Missouri Police Department by Chief Bernard Brannon, and continues to this day to be a controversial issue in many other departments. This innovation, along with his strong advocacy of an increased educational level for police officers, earned Chief Brannon a national reputation in law enforcement.

When Clarence M. Kelley became Chief of Police in 1961 we were a relatively modern police department, by traditional standards. For the most part, the officers were well trained, by contemporary standards, and dedicated to good police performance. Internally, there were problems. The previous Chief of Police and several high ranking officers had recently been indicted by a grand jury on matters of a malfeasance nature. The major issue of the Chief's indictment concerned inaccurate crime reporting and statistics. There were a number of cliques within the organization and this exerted a stronger influence on promotions and assignments than did objectively assessed merit and qualification.

At the time Clarence Kelley became Chief in 1961 he retired from the Federal Bureau of Investigation after more than twenty years of service. Though Kansas City was his hometown, he had not resided there for many years and was not at all familiar with the police department. There were many members of the department, especially among the top ranks, who resented an outsider being appointed as Chief of Police, as opposed to the position being filled from within the organization. Due to this fact and to the politics of the internal cliques, Chief Kelley immediately experienced difficulty in eliciting the candor and dependable information necessary for him to become acquainted with the department and its problems. He, therefore, reorganized the structure of the department, creating eight separate divisions whose commanders reported directly to him. While this is an unconventional structure and a very wide span of control for a police administrator, it served its intended purpose. It enabled Chief Kelley to break up the power cliques, to become familiar with the various operations in a first hand manner, and to assess the strengths and weaknesses of command and supervisory personnel. After he had accomplished these things he returned to a more conventional organizational structure.

Chief Kelley spent the first several years changing the climate within the department. He stressed the importance, in fact necessity, of integrity in both the individual and organizational sense. He recognized that no one person can administer such a complex function and organization alone, and he stressed the necessity and benefits of involvement of his personnel in the management and progress of the department. He believed and explained that there are two types of mistakes; mistakes of judgement and mistakes of the heart. 7
Assurance was given that honest mistakes in judgement made in the process of trying to do a good job would not negatively affect one's standing and future in the department. He changed promotional procedures so that promotions were based on competition and merit, and promotion by virtue of internal politics or favoritism was no longer possible. While the department was a good one by traditional standards when Chief Kelley took office, he was convinced that he had been given a mandate to make it a better one, the best one possible. While he undoubtedly quickly recognized some of the changes which were needed, he realized that constructive change cannot be forced and be successful, hence his efforts to change the climate of the organization to one of integrity, operational ethics, and involvement. The type of research, planning, and progress noted in the following pages could not have occurred had this climate not been created.

Chief Kelley also strongly believed in the utilization of technology in law enforcement and was responsible for the implementation of helicopters as another dimension of patrol. Also, through his efforts a computer was acquired, with the top priority application being that of service and assistance to the police officer on the street. Today the Kansas City, Missouri Police Department computer system serves over fifty criminal justice agencies in Western Missouri and Eastern Kansas in addition to our own, and has been termed by many as the best police computer system in existence.

While the department is under state control, as described previously, the operating budget must be appropriated by the city government. Like most other organizations, our financial needs have increased each year for many years. These increased needs have been due to a combination of economic inflation, increasing demands for quantity of police service, and the costs associated with programs to improve the quality of police service. In the face of these escalating budget requests, the City Council in 1965 insisted they be given an independent view of the department's operation.

The Board of Police Commissioners and chief Kelley readily agreed to this and as a result a contract was negotiated with the Public Administration Service of Chicago, Illinois for a study of the department. The study was very comprehensive in scope, including administrative, management, and operational facets.

The personnel complement of the department's Planning and Research Unit was increased for the purpose of working with Public Administration Service on the study. Their functions were to assist in acquiring and compiling requested information, provide liaison with the various organizational elements of the department, etc. The Planning and Research Unit had minimal involvement in determining the thrust of the study or in formulating the recommendations that would be forthcoming.

8

It was decided at the outset of the study that PAS would submit recommendations for change as they were formulated, and that if the change recommended seemed reasonable and held potential for improvement the department would proceed with implementation immediately, as opposed to deferring any and all changes until completion of the study. One reason for this was to get the consultants involved in implementation while still on site. The Planning and Research Unit also assisted in the implementation phase, mainly in a supportive or facilitative role, such as writing procedure manuals, etc.

Overall, the study resulted in a number of recommendations and changes throughout the department. Some of these changes have survived to the present, either in original or subsequently revised form, and others were totally unsuccessful and have long since been discarded.

One of the more significant and controversial changes concerned the organization of the patrol function. Prior to the study, command of the patrol function was vested in the commander of each patrol station area or patrol district, who was responsible to the commander of the Patrol Bureau. The station or district commander had twenty-four hour responsibility for his geographical area. Each had a subordinate field commander responsible to him for each of the three eight hour watches. This was changed to a watch-zone concept as recommended by PAS. There were three watch commanders, one on each eight hour shift, responsible to the Patrol Bureau Commander. Each watch commander was responsible for the patrol function for the entire city, but only during his assigned eight hour watch. The city was divided into three geographical zones, each having a zone commander responsible to his respective watch commander. Under this organizational structure, there was no one below the Patrol Bureau Commander who had twenty-four hour responsibility for the patrol function in any given area of the city. The command structure was built on an eight hour segment of the clock as opposed to a geographical area. Great difficulties were experienced with internal communications, transmittal of orders, citizen satisfaction, and personnel morale. Operation under this structure was almost totally unsatisfactory and in 1971, four and one-half years later, the department reverted to the previous command and organizational structure within the patrol function.

There were some worthwhile improvements and progress made as the result of changes made in response to recommendations made by PAS. Implementation of the changes and realization of the progress did not come easily, however. Hindsight makes it clear that the main obstacles encountered were due to the fact that personnel directly affected by the changes had very little input as to what those changes should be. There was resentment that "outside experts" could come into the department and tell us how we should do things. When people involved in an operation have the opportunity to be significantly involved in the identification of their own problems and development of their solutions they have a vested interest and intense commitment to successful implementation of those solutions. The total realization of this fact is perhaps the most valuable result of the PAS study, for the Kansas City, Missouri Police Department, and you will see that it was certainly kept in mind as we structured subsequent research and planning programs.

By the time the PAS study was over, the personnel complement of the department's Planning and Research Unit was approximately twenty. While it was originally intended that most of those transferred to the unit were there on a temporary assignment to work with PAS, the size of the unit was never decreased. Even though one of the PAS recommendations was for the continued existence and utilization of such a staff unit, the main reason the unit was not diminished in size or importance was that Chief Kelley strongly believed in its value to the continued progress of the department. He was convinced that intelligent decisions required that the problem or issue be accurately identified and described, that all pertinent information be accumulated, and that alternatives be identified and evaluated. Certainly he did not have the time to apply this process personally to all issues confronting him, so many of them were assigned to the Planning and Research Unit with a request for study and recommendation.

All members of the unit were sworn law enforcement personnel, with the exception of clerical personnel. The officers assigned to the unit were selected on criteria which emphasized past job performance, intelligence, commitment to professional excellence, and interest in the assignment. The assigned officers had practically no formal

training or experience in formal research or planning and they learned and improved through experience. Motivation was very high due to the challenge and to the firm knowledge that the Chief of Police sincerely attached great importance to the worth of the unit's product. Unit personnel proposed to Chief Kelley that he meet with them periodically for informal discussion of matters of current interest and concern. They felt such discussions would be very beneficial by permitting them to become exposed to his philosophy and goals on policing and department administration. He agreed to this, but at the very first such meeting made a statement to the following effect: "I can see where your feeling that an understanding of my personal philosophy and goals will be of assistance in your work, but I want to make one point very clear. I don't want you to ever give me staff work or recommendations which is merely an attempt to give me what you think I want to hear. If you do that your contribution to this department will be of minimal value. I want you to approach all issues objectively and give me the benefit of *your* best thinking and *your* recommendations. It is my responsibility to accept or reject your recommendations, and in so doing, I am totally responsible for the results, good or bad."

Within the department, the Planning and Research Unit was frequently referred to as "the ivory tower bunch," "the empty holster crowd," and similar terms, sometimes seriously and sometimes in jest. Conscious efforts were made by unit personnel to consult with those assigned to functions potentially affected by the project being worked on and therefore, department personnel were more receptive to change resulting from such internal staff work than had it been developed by outsiders. This does not mean that only information and opinions from within the department were gathered or considered. Depending on the nature of the project, input was also sought from other police departments, criminal justice agencies, business, industry, etc., i.e., any source deemed appropriate and pertinent.

The Planning and Research Unit was kept busy with staff studies and development concerning matters of current and pressing urgency. There was a desire and recognized need by both department management and the staff of the unit to become involved in some research and planning of a more long range nature, but it seemed that the time was just not available. Prompted by these circumstances, plus the recognition of the potential beneifts of involving a greater number of departmental personnel in planning, personnel of the Planning and Research Unit in October, 1969, proposed to Chief Kelley the formation of several task forces.

It was proposed that each task force include representation from command, supervisory, and patrolmen levels and that they be charged to research and submit recommendations for future direction relative to some rather broad and general subject areas. This general concept was discussed with Chief Kelley and he reacted with wholehearted support. As a first step he requested that each commanding officer in the department submit a paper to him discussing their assessment of the strengths and weaknesses of the department and their ideas for future changes and direction for the department in pursuit of increased professional excellence. Not only was this a first step in the intended task force organization, but the responses were of great value to Chief Kelley in helping him further assess the individual strengths, weaknesses, and potentials of his commanding officers. Following Chief Kelley's review of these papers they were given to Planning and Research Unit for review, summarization, and identification of the subjects receiving significant attention. In May, 1970, eight task forces were formed and each was charged to address themselves to one of the following subject areas: (1) regionalization of certain police functions; (2) possible additional sources of revenue for the operating budget; (3) educational standards for police; (4) supervisory training and development; (5) human relations, both within the department and with the community; (6) improvement of investigative procedures; (7) improved patrol concepts and procedures; and (8) improved inservice training programs. Each task force was composed of two commanding officers, two sergeants, and two patrolmen or detectives. The commanding officers were appointed by Chief Kelley and they then selected and recruited the other four members of their respective task forces. Since the department was very undermanned, it was necessary to require that all task force members continue their primary duties full time and address their task force assignments as time permitted. They were told that they were free to seek information and assistance from any source willing to provide it, but there was no money available to hire consultants or staff. From this point on, they were on their own except for what assistance the Planning and Research Unit could provide relative to possible sources of information and staff study methodology.

The reports received from these task forces ranged all the way from very brief, elementary and superficial, to very comprehensive with much effort and good thinking quite obvious. Some of the reports received no further action or attention once they were read due to their lack of substance and/or a lack of the means and resources to pursue the subject at the time. Some resulted in varying degrees of changes and new programs within the department in the following two years. Those which prompted change or new programs concerned supervisory and executive training and

development, human relations, and in-service training.

Overall, the quality of efforts exerted and reports submitted were quite commendable when one considers the circumstances under which the task force members were asked to produce. They did not possess formal knowledge or skills in research, problem identification, or program development and they were not provided funds to avail themselves of assistance in these areas. They were expected to continue performing their normally assigned duties and do their research and produce their report as an extra assignment. Most of the task forces were composed of members from various units of assignment and, in some instances, who worked different duty hours. While it was originally felt that such diverse representation within a task force would be a positive factor, hindsight indicates that it was not. It is difficult for a person to grasp, get highly motivated toward, and pursue issues foreign to his experience and assigned duties. It also made it very difficult to schedule task force meetings. Another aspect which presented problems was that most of the assigned subject areas were too general and broad and there was much floundering in attempts to identify specific and definable issues to pursue.

Probably the most significant benefits derived from this task force program was the experience and effects on those who were members of the task forces, and not specific changes resulting from the reports. It emphasized the sincerity of Chief Kelley's philosophy of participatory management and desire for the thinking of all members of the department; it stimulated conceptual thinking; and it expanded the participant's awareness and understanding of problems and issues confronting law enforcement beyond those of the individual's specific normal duty assignment.

The next significant phase of the department's research and planning experience resulted from the combination of two events, the creation and mission of the Police Foundation and approval by the voters in Kansas City of an increase in the city's earnings tax from .5% to 1%.

The Police Foundation was created in 1970 with a five year, 30 million dollar grant from the Ford Foundation and a mandate to "assist police agencies in realizing their full potential by developing and funding promising programs of innovation." Representatives of the Foundation visited a number of major police departments to become more familiar with current policing methods and problems and to try to assess the capacity of the departments for the development and implementation of innovative programs. The Kansas City, Missouri Police Department received such a visit by three representatives of the Foundation in early 1971. In the summer of 1971, the Foundation sponsored a two week conference at the University of Wisconsin, attended by members of the departments which had been visited: New York, New York; Baltimore, Maryland; Cincinnati, Ohio; Detroit, Michigan; Dallas, Texas; and Kansas City, Missouri. This conference involved discussion of policing problems, programs, and potentials and was attended by Chief Kelley and six commanding officers from Kansas City. The Foundation had indicated that following the visits to the departments and the conference they would select several of the departments and award them major grants. Shortly after the conference it was announced that the Cincinnati, Ohio and Dallas, Texas Police Departments would receive grants. Since it was very unclear what the potential grants would be for or what relationship the Foundation expected to establish with the departments, the Kansas City, Missouri Police Department did not pursue the award of one of these grants. It is not clear what consideration on the part of the Foundation resulted in Kansas City not being offered one of the grants.

In December, 1970, the voters of Kansas City, Missouri approved an increase in the city's earnings tax from .5% to 1%. The city government had made a commitment to the voters that the great majority of the resulting revenue would be spent for public safety, including the addition of 350 officers to the police department. The department actively and vigorously campaigned for passage of the earnings tax increase, promising that 280 of the 350 additional officers would be assigned to patrol and specifying how many were to be assigned to each patrol division so that voters would know what to expect in the way of increased visible police protection in their particular areas of the city.

Chief Kelley recognized that the addition of these officers provided a rare opportunity to reassess existing patrol strategies and procedures and to develop plans for the deployment and utilization of the additional officers in the most beneficial manner possible. In fact, he felt we were ethically obligated to do so. In late August, 1971, Chief Kelley and members of the command staff again met with representatives of the Police Foundation. The Foundation was informed of the department's intent to study patrol strategies and problems and to pursue improvement and they were invited to consider joining with us and assisting us in these efforts. We made it very clear that any projects were to be a department venture not a Foundation venture; that we would insist on retaining control and responsibility for what was done. Within this context we took the position that we would appreciate the assistance the Foundation could provide and would make all possible efforts to work with them.

After lengthly discussions the Foundation agreed to join with us. While the department fully

intended to embark on these efforts concerning patrol, with or without the assistance of the Foundation, we had practically no resources for consultant assistance or other expenses, therefore, the assistance of the Foundation was of great value and facilitated much more comprehensive efforts and projects. The Foundation provided funds for such assistance as consultants, travel to other departments to study various programs, rental and/or remodeling costs for office space, overtime pay, clerical staff, and evaluation.

In October, 1971, four task forces were formed, one within each of the three patrol divisions and one in the Special Operations Division. This division is composed of patrol support functions, i.e., Tactical Unit, Helicopter Unit, Canine Unit, and Police Reserve Unit. Each task force was composed of six to eight members: the division commander (rank of Major); one captain; and the remaining members with the rank of sergeant and patrolman. All three watches, or shifts, were represented in each of the patrol division task forces. The division commander was chairman of the task force but each member had equal input and vote without regard to rank. To provide process assistance and support in problem identification, research, and program development, one officer of the Planning and Research Unit and one Police Foundation consultant was assigned to work with each task force.

Each task force was given a mandate to identify the most critical problems confronting its respective division and to develop and submit recommendation for addressing them. Chief Kelley assured his total support and assured the task forces that their recommendations were to be submitted to him, that he would thoroughly study and consider them, and that he would make the final determination as to their implementation. He stressed the absolute necessity for integrity in all that they might do.

The task force approach was chosen for three main reasons: (1) involvement of people affected most by a program in the development of that program greatly increases the commitment to implementation and enhances the success of the program; (2) it was believed that the persons working in the divisions could best and most accurately identify and assess the contemporary problems facing their respective division; and (3) a firm belief in the individual and collective capacity of the patrol officers. While the task force approach is not usually the most expeditious and efficient procedurally, it was believed that the value of (1) and (2) above made this approach much preferable to any other alternative. In organizing and setting up the task forces we tried to apply lessons learned as the result of mistakes made with the task forces created in 1970 and previously described herein.

It was intended that the task forces be, to the extent possible, representative divisions, and they were urged to develop and maintain the best communications possible in order to receive input from all personnel and to keep them informed of what was going on. This was not an easy thing to do, especially during the early stages when the task forces were involved in general discussions and attempting to define their direction. The various methods used in attempts to establish and maintain communications included inviting division personnel to attend task force meetings, memorandums, having task force members attend regular roll calls periodically, and having a task force member ride patrol with the officers.

Task force activity began initially with periodic meetings, usually weekly, and members otherwise continuing to perform their normally assigned duties. A number of trips to other cities were taken by task force members to study other patrol operations and programs. When this occurred, the member(s) making the trip were relieved from their normal assignment and were considered to be on temporary special duty status. Later in the process some members of the task forces were relieved of normal duty and assigned full time task force duty to pursue program development.

Shortly after the task forces were formed a Task Force Coordinating Council was created. The council was chaired by the commander of the Patrol Bureau and included the commander of each of the divisions having task forces and the commander of the Planning and Research Unit. The purposes of this council were to provide coordination between the task forces, exchange information of common interest, avoid unnecessary duplication of research and other efforts, keep the Patrol Bureau commander informed of task force activity in all of his divisions, address policy issues raised by task force activities, and review task force program proposals. As previously noted, Chief Kelley retained the responsibility for final approval or disapproval of task force proposals so the council could only attach their recommendations for the Chief's consideration.

At the inception of the task forces the consultant assistance provided to each task force by the Police Foundation consisted of individuals with primary employment and responsibilities elsewhere in the country. These persons would fly in to Kansas City for task force meetings, usually for one day per week. It soon became evident that this was not a satisfactory arrangement and would become even less satisfactory as the task forces got closer to program development and implementation. The task forces felt that the arrangement did not permit the degree of involvement and commitment on the part of the consultant which they felt was necessary and that the limited access to him was not adequate for their needs.

As a result of the dissatisfaction with the "fly-in fly-out" consultant arrangement, the Operations Resource Unit was created as an organizational element of the Patrol Bureau. Persons with needed skills were hired by the department on a contract basis with funds provided by the Police Foundation. The unit was headed by a regular department member. By this time it had also been recognized that, in addition to whatever might result from the task force programs, one of the potential benefits of the relationship with and support of the Police Foundation was the acquisition or development of research and program development skills within our department, which would remain with us and be of value beyond the current task force program and affiliation with the Foundation. Accordingly three patrolmen with a high interest and potential for this type of work were selected and transferred to the Operations Resource Unit. This unit did not have the role or authority for making significant decisions; their primary purpose was providing process support to the task forces. In addition to the activity of direct and active process support, the unit provided computer programming capacity, compiled a library of programs of interest on a national scope, catalogued information emerging from task force activities, and provided access to consultants available nationally when needed.

12

All of the task forces successfully completed the process of identifying problems, prioritizing them, and selecting specific problems for which they developed and implemented new programs or experimental research. Several of these programs, after trial and evaluation within the division of origin have been implemented and institutionalized throughout all patrol divisions. Purpose and space of this paper do not permit a discussion of each of the projects. One, the South Patrol Division project, will be briefly described because it was experimental research in nature, was of great significance to the field of policing, and demonstrated that a police organization can design and conduct meaningful research.

In response to instructions to all of the task forces to identify the most critical problems confronting their respective divisions, the South Patrol Division Task Force identified five problem areas: (1) residence burglaries; (2) juvenile offenders; (3) citizen fear of crime; (4) public education about the police role; and (5) police-community relations.

> "Like the other task forces, the South Task Force was confronted next with developing workable remedial strategies. And here the task force met with what at first seemed an insurmountable barrier. It was evident that concentration by the South Patrol Division on the five problem areas would cut deeply into the time spent by its officers on preventive patrol. At this point, a significant thing happened. Some of the members of the South Task Force questioned whether routine preventive patrol was effective, what police officers did while on preventive patrol duty, and what effect police visibility had on the community's feelings of security.
>
> Out of these discussions came the proposal to conduct an experiment which would test the true impact of routine preventive patrol. . . .
>
> As would be expected, considerable controversy surrounded the experiment, with the central question being whether long-range benefits out-weighed short-term risks. The principal short-term risk was seen as the possibility that crime would increase drastically in the reactive beats; some officers felt the experiment would be tampering with citizen's lives and property.
>
> The police officers expressing such reservations were no different from their counterparts in other departments. They tended to view patrol as one of the most important functions of policing, and in terms of time allocated, they felt that preventive patrol ranked on a par with investigating crimes and rendering assistance in emergencies. While some admitted that preventive patrol was probably less effective in preventing crime and more productive in enhancing citizen feelings of security, others insisted that the activities involved in preventive patrol (car, pedestrian and building checks) were instrumental in the capture of criminals and, through the police visibility associated with such activities, in the deterrence of crime. While there were ambiguities in these attitudes toward patrol and its effectiveness all agreed it was a primary function." [1]

Out of these discussions came a task force proposal to conduct an experiment to assess the value of the traditional routine preventive patrol. Chief Kelley, displaying a great degree of administrative courage when one considers the strong tradition being questioned and the unknown outcome, granted his approval to proceed with the experiment. In doing so he imposed two constraints: (1) the department's responsibility to serve and protect the public must not be neglected; and (2) the department's normally low response time to calls for service must not be impaired. It was agreed that crime statistics would be monitored closely on a weekly basis and that any significant increase in the experimental area would result in prompt termination of the experiment.

The experiment was conducted in a 32 square mile area of the South Patrol Division having a

[1] George Kelling et al., *The Kansas City Preventive Patrol Experiment A Summary Report* (Washington, D.C., 1974), p. 7–8.

1970 census population of 148,395. The 15 patrol beats in this area were computer matched on the basis of crime data, number of calls for police service, ethnic composition, median income and transiency of population into five groups of three. Within each group of three beats one beat was designated as reactive, one as proactive, and one as control. In reactive beats all routine preventive patrol was withdrawn. The assigned patrol unit responded to and handled all calls for service but when not so dispatched and occupied remained on the beat perimeter or patrolled in an adjacent proactive beat. In the proactive beats the level of routine preventive patrol was increased from two to three times normal through the assignment of additional patrol units and patrolling of reactive units. The level of patrol in the control beats remained normal, with one unit assigned to each beat patrolling in normal manner.

The experiment was initially started on July 19, 1972, but was suspended in mid-August when it was recognized that experimental conditions were not being adequately maintained and that some problems were evident. Necessary revisions were made in instructions and guidelines and the experiment was resumed on October 1, 1972, and reached a successful conclusion on September 30, 1973. Data was collected by means of ten different surveys and questionnaires, interviews, observers riding with officers, and from departmental data (crime, traffic, arrest, dispatch, officer activity, and personnel records).

The public was aware that an experiment was being conducted but was not informed of the exact nature of police patrol presence in the various beats nor specific locations of the beats. In one incident a businessman was informed by an opponent of the experiment that his business was located in an area from which all police patrol had been withdrawn and a protest was expressed. Chief Kelley met with business representatives of the area and explained the nature and purpose of the experiment and that it was being closely monitored. At the conclusion of his explanation he received a standing ovation from those present.

The results of the experiment disclosed that the varying levels of routine preventive patrol had no effect on actual crime, reported crime, community attitudes toward police on delivery of police service, response time, or traffic accidents. Of 648 individual statistical comparisons made to produce the major findings, statistical significance occurred only 40 times.

In July, 1973, Chief Kelley became Director of the Federal Bureau of Investigation and in November, 1973, Joseph D. McNamara became Chief of the Kansas City, Missouri Police Department. Chief McNamara quickly expressed his support of the department's research orientation and efforts to improve our reputation as one of the best police departments in the nation.

At the present time, the department is involved in three very significant projects.

1. Directed Patrol—This project was implemented in the East Patrol Division on July 1, 1976, and is a natural follow-up to the South Patrol Division Preventive Patrol experiment described above. Given the results of the South experiment we felt obligated to develop more productive methods of utilizing the uncommitted time of patrol officers. One problem in doing so is the fragmentation of such time. The Directed Patrol program, developed by an East Patrol Division Task Force, has two major components. The first seeks to assess priority of calls for service, with some responses being delayed, some citizens being requested to come to the station to make reports, and some reports being taken by phone. This is an effort to realize uncommitted time of patrol officers in larger and more predictable time increments so that it can be utilized in planned and directed patrol activity. The second component involves the utilization of that time in various programs directed toward crime prevention. Financial support for the development of the project was provided by the Police Foundation and funding for implementation and evaluation is from an LEAA grant. 13
2. Response Time Analysis Study—Police response time has long been assumed to be a very critical factor in police patrol effectiveness, especially with regard to apprehension of criminal offenders. A number of studies have previously been conducted, but none of sufficient scope and quality to prove or disapprove traditional assumptions. This study is a very comprehensive and sophisticated project started on October 1,1 973. The continuum from crime or other police incident occurrence to contact between the responding officer and the citizen is being measured in minute intervals for the purpose of assessing the effects of variable response times on arrests, witness availability, victim injury, and citizen satisfaction. A secondary objective is the analysis of problems and patterns of citizens reporting crime. This study is funded by the National Institute of Law Enforcement.
3. Domestic Violence—One of the many traditional assumptions in law enforcement is that the police are powerless to have any preventive effect on homicides and aggravated assaults because most of them occur between relatives or acquaintances, many are spontaneous, and/or most occur inside

buildings or other locations not visible to police patrol. In 1972 a sergeant assigned to our Planning and Research Unit gathered and analyzed a large amount of data from police reports of homicides and aggravated assaults, arrest records, and dispatch records. He concentrated on those of a domestic nature, which account for a major portion of the homicides and assaults. He found that in the two years preceding the offenses in a domestic setting the police had contact with either the victim or suspect, or both, in responding to and handling disturbance calls. In 85% of these cases the police had at least one such previous contact and in 50% of the cases we had five or more such contacts. Is there something the police can do in these contacts to forestall a future homicide or aggravated assault? The East Patrol Division recorded data on numerous variables observed in the process of handling disturbance calls. There is a very strong indication that various interacting variables can provide some ability to predict potential for future violence between the participants of a domestic disturbance. If this is true, it is felt that the police can refer such people to an appropriate social service agency for assistance, thereby reducing the incidence of domestic homicides and assaults. In July of this year, the National Institute of Mental Health awarded a grant to the department for further analysis of the data collected and the collection of additional data.

14

It might seem to the reader of this paper that what has transpired in the Kansas City, Missouri Police Department insofar as research, experimentation and planning resulted from a grandiose master plan or schedule developed years ago. Such is certainly not the case. To a large extent our efforts and progress have been reaction to contemporary events and opportunities. One thing that was deliberate, and I'm sure planned, was the creation by Chief Kelley of a climate within the department which encouraged involvement and innovation. Sincere and strong top management support for such is absolutely essential to meaningful and successful efforts such as have been discussed. Along with this strong support, management must assume a facilitative role as opposed to a strong directive role; an overly directive role stifles initiative and participation of personnel within the organization. All of our patrol task forces were initiated at the same time in 1971. One of these task forces struggled much harder and took much longer than the others to "get off the ground" and start making some meaningful progress. There is general agreement among those who monitored the process that this was due to the fact that the commander of that division was quite authoritarian in his personality and management style.

The department has recognized and realized many benefits and advantages of the task force approach. Some of the very significant ones are:

1. It provides an environment for personnel development and enhances capacity to properly handle discretion.
2. It provides an opportunity to identify highly competent personnel at all levels of the organization.
3. It increases communication, coordination, and morale within the organization. Prior to the patrol task forces there were frequent requests for transfers to other parts of the organization. As the task forces got more involved these requests for transfer out of patrol decreased drastically and, in fact, we started receiving requests for transfer to patrol from other elements.
4. It improves the ease and success of implementation of change due to the involvement and vested interest of those affected by the change. Consider the statements of one of our officers who was involved in one of the patrol task forces:

   "They've said policemen fight change. Well, that may not be true. It may have been the method of change, rather than the change itself, that was resented. The patrolman wants change but he wants to have a part in deciding what that change will be."

There is no intention to create the impression that the task force approach is appropriate for all circumstances or that it does not have negative aspects. It is a slow and time consuming process and increases the difficulties in controlling variables during the evaluation phase of a project. We have also found it not to be the best approach for very technical areas or issues not a part of the everyday duties of the task force members.

Some keys to successful operational research

Based on my observations of our experiences in the Kansas City, Missouri Police Department over the past decade, there are several very key points in conducting worthwhile and successful operational research.

The first thing which must exist is top management support and commitment to such efforts and programs. Without this it would be totally futile to try even the first step. This factor has been discussed in some detail in the preceding pages of this paper, but its importance cannot be overemphasized.

Another very important consideration is the meaningful involvement of the personnel of the organization including, in fact especially, those at

the rank and file level. Again, this has previously been discussed and stressed in preceding pages, but bears repeating. Too many managers are inclined to believe that the people of an organization are totally against any changes, except increases in the pay check and decreases in working hours, and that they will do all within their power to resist change. That just is not so. They do like to play a part in their destiny and it is to the organization's benefit to let them do so. Of course, there will be individuals who are exceptions but it has been our experience that the enthusiasm and satisfaction generated within the majority results in peer influences preventing those individuals from generating serious or successful resistance. It should go without saying that the reason for and subject of any research project or program must be legitimate and have as its goal the improvement of the organization and the service it provides. Research purely for the sake of research should be taboo. If a manager cannot project the justifications and potential benefits in a totally convincing manner it must be questioned as to whether the project or program is warranted.

Total honesty with the personnel of the organization is a must. They must be truthfully informed of the purpose of the research and the methods to be employed. I am aware of one organization which utilized field observers to gather data for their research. The rank and file were given a ficticious account of what type of information the observers were to gather. Once this deception became known the ability to collect accurate and reliable data in that organization ceased to exist. Even if the rank and file members are not an important source of data for the research or are not otherwise involved in the process, a lack of factual information will likely result in rumors and inaccurate perceptions, thereby detracting from the value and success of the research. In our department we utilized various means in efforts to keep personnel informed. Personnel directly involved in the projects were urged to utilize every opportunity to communicate with their peers, briefings on current projects were included in recruit and in-service training classes, articles were printed in the department newspaper, memorandums were written and distributed, and projects were discussed in staff meetings. It takes a lot of effort to keep information flowing to all parts of a large organization but the dividends make those efforts worthwhile, in fact necessary.

Operational research within a public service agency does present problems which are not as likely to be encountered in a product producing organization or one whose service is less essential and visible. We must be continually responsive to the publics' needs and demands for our service, often times on an unpredictable and emergency basis. The research must be conducted in such a manner that our ability for such response is not compromised. The ground rules for assuring this must be set forth at the beginning, and the research must be designed and structured with full understanding and consideration of these rules. This initial effort can avert many of the problems that would be encountered, but there is no way to anticipate all problems relative to conflict between the project and its evaluation and what would be otherwise normal changes such as personnel reassignments, changes in personnel deployment, changes in organizational structure, changes in tactical strategies, etc. When these conflicts arise those with primary responsibility for project administration and those responsible for operations in provision of the agency's daily service to the public must confer and collaborate in resolving the conflict in the proper and best interest of the public. This is not as easily done as said but it is necessary and possible. The Kansas City Police Department has certainly encountered some very knotty problems of this type and a gentleman who will speak to you, Dr. George Kelling who was on the Police Foundation evaluation staff for some of our projects, can relate the details of some of those problems and their outcomes better than this writer.

Evaluation results and decision-making: the need for program evaluation

Lee Sechrest
Professor of Psychology
Florida State University
Tallahassee, Florida

16

This paper attempts to make the strongest possible case for systematic evaluation of programs and other interventions directed toward the resolution of operational problems in service agencies. It is based on the premise that many administrators have not thought through their own needs for information and the role that research data can play in effective decision making.

Making decisions in any complex, real-life setting is never a unidimensional, or even a simple, process. In order to make adequate decisions the wise executive knows that it is necessary to have good information on the effectiveness of some proposed act or intervention. For example, before deciding whether to buy a certain type of emergency vehicle, a wise executive would want to know whether the vehicle could do what it was designed to do, whether it was engineered in such a way as not to create more problems than is solved, whether it might also produce some nonobvious benefits by making possible the performance of other important tasks, and he would want to know whether the vehicle was really the best of its type. All the above established in the affirmative, the decision to purchase the vehicle should not automatically be made. Other factors of equal, and perhaps greater importance, would have to be considered. First, economics would be important. The cost of the vehicle would be important, and maybe critical. No matter how good it was, an emergency vehicle might be beyond the budget even imaginably available to the community, and even if affordable, the vehicle might cost too much more than the closest competitor. Practicalities might also be important if it appeared that delivery of the top-rated vehicle might be long delayed or if service might be unduly difficult. Political considerations might arise. Suppose the emergency vehicle in question were manufactured in the U.S.S.R.? No one would dare recommend its purchase. But even if it were only manufactured in another state and had to compete with a locally manufactured product, it might be politically unfeasible to recommend its purchase.

The complexities no more than hinted at above are severe enough for the fairly ordinary affairs of public institutions, e.g., purchase and cleaning supplies, revision of accounting systems, deciding whether to stagger times of work shifts, but they are increased almost immeasurably when decisions have to be made in the context of ongoing and critical public services. To revert to the example noted earlier, if the decision were which model of a garbage truck to buy, the fact that one model might result in slightly higher spillage than another would be troublesome but scarcely beyond dealing with. When the problem, however, is the purchase of emergency vehicles and the issue is the saving, or possible saving of lives, feelings run high and decisions must take more factors into account. It follows, then that decisions in critical public services may not reflect quite so clearly the harder more factual information on effectiveness of a proposed intervention.

The position taken here is that despite the complexity of decision processes in such areas as emergency medical systems—and, as we shall see, police systems also—data on effectiveness based on careful evaluations is still an important element in the decision process, even if the final decision goes against evaluation results. An administrator may find that a suggested change in operations would be economically unfeasible, that it would be politically unacceptable in his community, that it would be resisted too strongly by employees at lower levels, and he might decide against implementing a change even though on other grounds it would be desirable. It is the contention of this writer that the administrator should know exactly what he is sacrificing, the price he is paying to maintain labor peace, to avoid having to ask for additional funding. There is absolutely no advantage in making decisions in which one of the important elements is an unknown. If, for example, a proposed new emergency vehicle would be little more effective than those already available and the other costs are sizeable, the administrator's decision is a simple one. If, on the other hand, the proposed vehicle would actually perform significantly better and result in better outcomes for emergency cases, the administrator can understand his decision as an honest and rational one and can also take comfort

in the knowledge that if some of the other factors change, e.g., economic situation improves, there is a good basis for reconsideration.

Therefore, we can only recommend that administrators cooperate in, indeed insist on, obtaining the best information possible about program effectiveness since that information is not only an important but critical element in managerial decision-making.

What is program evaluation?

At this point it might help to make clear just what is meant by a program or an intervention, what is meant by an evaluation, and what is meant by effectiveness. In the broadest sense we mean by a program or an intervention, *any* alteration in an organization, including changes in personnel, in equipment, or in operating procedures, and that is intended to improve the operations of the organization and make it more likely to achieve its goals. When a rescue squad purchases a new communications system, when a department of public safety replaces an administrator judged ineffective, when an emergency services delivery program is regionalized, when all rescue team members are required to undergo some training program, these are all instances of interventions of the type we have in mind. Then when we say they should be evaluated we mean that some process should be established to determine whether the intended effects are achieved. If a baseball team in a slump fires its manager, it is reasonable to keep track of performances of individual players and of the team as a whole. If a new communications system is purchased, then procedures should be set up to determine whether communications are affected. Does the delivery of emergency services change following regionalization? Do trained ambulance attendants perform differently as a result of their training? Then by effectiveness we mean whether the change(s) is in the intended direction, whether the change is about as large as was anticipated, and whether there are unexpected additional benefits for disadvantages resulting from the intervention. A new rescue vehicle might not only be medically more desirable but might improve morale and pride of the squad. A new administrator might produce greater efficiency in operations but also produce undesirable turnover in personnel over the long run.

What we are recommending is that *all* changes should be considered to be temporary, to be experimental, and that procedures should be established to evaluate their effects. Perhaps that may seem an unrealistic recommendation, but in our view to do less is irrational. There is not much purpose in replacing one administrator by another in order to improve organizational performance without having some way of knowing whether the improvement takes place. It does not make much sense to buy a new piece of equipment without having some plan for determining whether it works better. It is obvious that different types of decisions may be evaluated in different ways. and not all require formal study and experimentation. Some evaluations occur in the normal course of events, and if the risks involved in simply waiting to see what happens are not too great, needed data will often emerge. There was a recent newspaper article reporting that steel-belted radial tires are undesirable for cars likely to be driven over 100 miles per hour because failure of heat dissipation leads to blowouts. This fact was discovered because of failure of such tires on police cars used in high speed chases. It does seem just possibly a bit unfortunate for a police department proudly outfitted with steel-belted radials on its cars to learn that such tires are not such good choices right in the middle of a high speed chase. Note that even in this case, however, the conclusion was made possible by accumulating data across a number of different jurisdictions. Think how long it might have taken for 100 on car police departments scattered around the country to learn the same thing. Obviously if a major decision is to be made, or if a decision is to be made which is not easily reversible, simply waiting to see what happens is weak evaluation strategy.

Some evaluations are pre-performed to at least some degree. Specifications for equipment, as an instance, are an attempt to ensure that the equipment will perform as expected. From a strictly hardware, technological standpoint, it may be possible to draw up and enforce specifications in advance. Even in some other areas technology may be sufficiently advanced that a change can be made with reasonable confidence of effectiveness. For example, not every training program has to be evaluated in every setting. Eventually one becomes confident that a given type of training is a desirable thing. However, there are good reasons for making conservative estimates of the probable effectiveness of new programs and for making at least some probing efforts to determine that the programs are having their desired effects.

It is tempting to think that at least some types of programs or other interventions can be *assumed* to be effective, e.g., on local grounds or by analogy. Based on reviews of many other programs and innovations in many other areas, we have concluded that it is risky, if not downright hazardous, to assume anything about the probable effect of a program. A large number of examples can be cited of programs and practices which were assumed to be desirable or which became standard practice before any evidence of effectiveness was available and which have not only in some instances proven worthless, but worse, have on occasion proven dangerous. It is also unfortunately the case that at least some of these programs persist and even pro-

liferate despite proven ineffectiveness. However, before pointing to specific examples, it might be noted also that even if a program can be assumed to be desirable, to be on the whole an improvement, it is much more difficult to know whether any assumed benefits are proportional to costs. It may be possible to demonstrate conclusively by purely technical evidence that a new communications system will result in reduced dispatching time, but if the system requires better trained personnel, renovation of space, etc., it may prove deceptively expensive. But even if all those factors are known, it may still be highly questionable whether the projected decrease in dispatch time will be worth the costs.

18

Wastefulness of ineffective solutions to problems.

The problem with ineffective "solutions" to problems is that they are wasteful, usually in several ways, and hence should not be tolerated. In these days of increasing pressures for accountability on the part of public institutions, it is going to be increasingly necessary to produce positive evidence of effectiveness of new programs and changes in old ones. Ineffective programs are, quite obviously, wasteful of resources: space, time, talent, money. The city of Miami Beach, Florida, mandates that a physician ride along on every emergency vehicle run. If that physician does not in some substantial degree improve the results of emergency runs, then money—a good bit of it—and talent that could well be used elsewhere are being wasted. However, at a less obvious level than the wasting of resources, ineffective programs are wasteful because they often involve substantial and important opportunity costs,. i.e., money or energy invested in one enterprise is not available for other, perhaps much more productive purposes. A relatively obvious opportunity cost is the economic one: purchase of one $13,000 emergency vehicle means that two $6,500 vehicles cannot be purchased. The hiring of a full-time emergency physician may mean that two fewer nurses can be employed. Money spent to install radiographic equipment in an emergency room will not be available to renovate space to improve work-flow.

It needs also to be recognized that ineffective programs may be worse than simply wasteful because they detract attention and energies from problems badly needing solution. For example, it has been noted that almost any anti-delinquency program, even if it is quite ineffective, reduces public anxieties about the problem and any resulting pressures for a solution. It has been argued that every ineffective delinquency program sets the field back about five years because that is how long it takes to discover that it is not working. The situation cannot be different in the health field generally and in emergency medical services delivery specifically. Think of the many changes in the EMS field that have been made with the promise, but not the demonstration, of effectiveness which have been or may now be called into question. And think how those very changes have retarded further explorations into the problems involved. We want to reiterate the point here that we believe it essential to plan for the best possible evaluation of every change, or innovation, or new program. We believe that absolutely nothing about effectiveness can be assumed.

Examples of unevaluated bad ideas.

Perhaps it might help at this point to give a few examples of how reason and logic have led to erroneous conclusions, sometimes with results that have been quite unfortunate. A good initial example, because it pertains to the training of personnel involved in delivery of critical public services is the set of assumptions that has long existed about appropriate training for police personnel. Since it is evident that police are often subjected to considerable stress, that they must cope with danger, harassment, enforced quasimilitary discipline, and the like, it has seemed evident to just about everyone that police training should prepare officers for those very experiences by providing occasions, preferably numerous, of a high degree of realism, on which they can practice the appropriate responses. Consequently police training has been militaristic, physically and emotionally demanding, marked by stern and stressful discipline, etc. A few years ago it occurred to H.H. Earle (1973), an officer in the Los Angeles County Sheriff's Department, that the assumptions on which so much of police training has been based just might be wrong. So he developed an alternative training program characterized by relaxed discipline, rational exercise of authority, minimization of artifically induced stress, and the like. Half of the recruit class were assigned randomly to the traditional training program and half to the new experimental program. The experimental program proved to produce patrolmen better in every respect, both at the conclusion of training and upon follow-up. The experimentally trained class were even judged later to wear their uniforms better, and they scored significantly better in marksmanship. Can anything about the training of EMTs or paramedics be taken for granted?

Over the years one of the convictions that has been prevalent about delinquent youth is that they come from rather generally disturbed families and that they need some sort of substitute parent, e.g., a "big brother," at least to tide them over, to help provide some of the attention and warmth that they fail to get at home. In the meantime, the family should receive some sort of therapy or counselling. A recently published study by the Institute for So-

cial Research at the University of Michigan suggests that not only are those assumptions not tenable, they may in part, be absolutely wrong. An experimental test of the "volunteer" delinquency worker program showed that it is, at best, of no value, and a further study showed that requiring the families of delinquents to participate in counselling programs was *worse* than leaving them alone (Berger & Gold, 1976).

The above are but two of *many* examples that could be adduced. Anti-drug abuse programs based on the very best of assumptions have proven generally worthless. The logic of probation and parole is inescapable, but neither seem to work at all. The state of Maine has recently, by action of the legislature, given up on parole altogether. When prisoners are released, they are released, and that is it. Although controversial, a recent report on the effectiveness of rehabilitative techniques with criminal offenders (Martinson, 1974) concludes that there is *no* rehabilitative power, however logical and appealing, that produces results in any dependable way.

The medical and health fields can provide as many, and equally good, examples. Cardiac Intensive Care Units may be of little or no value and even harmful in some cases. Coronary artery bypass surgery is quite logical and, on the evidence, little justified. Health Maintenance Organizations are proliferating around the country because they seem like a very good idea. There is as yet no evidence of their effectiveness and some modest pieces of evidence suggesting that they may be of little value. Health education is clearly a good idea, but at least as it has been implemented, it is a waste of money and effort. An interesting note on health education comes from Victor Weingarten, President of the Institute for Public Affairs, who found that five major voluntary health agencies were spending more than $100 million per year for health information programs. Yet over a period of ten years there were only two instances of any attempt by any of the agencies to evaluate any of the material. An insurance company spending $2 million per year for health information has never had an evaluation of the materials over a period of 20 years (Weingarten, 1974). A great deal of money and effort is being invested in the development of PSROs with almost no evidence at all that they will have their intended effects and with distinct risks that they will have quite undesirable side effects.

Two examples involving monetary considerations are of special interest. New York Bell Telephone Company concluded that they were spending too much money providing information services to subscribers who ought to look up the numbers in the telephone directories. They calculated that by instituting a charge for information service, which involved a commitment to refund $.30 to every subscriber not using information service, the company could save a great deal of money. However, subsequent to the invoking of the information service charge, there was such an enormous increase in requests for directories accompanied by unanticipated costs in refunding the $.30 to the huge number of subscribers who proved not to use information, that the company was faced with a very sizeable net loss, a figure around $2 million. A relatively small scale experiment might well have suggested what did in fact happen. Another example involving money is the hospital precertification program which was supposed to save Medicare and Medicaid funds by providing assurance that every hospital admission is, in fact, medically justified. However, precertification involves costs, and Drs. Thomas Bice and David Salkever are currently analyzing data which suggest that the "certificate of need" in fact resulted in a net *increase* in costs of hospitalization, probably by about $5.00 per hospitalization. (Bice personal communication). Not much, but when aggregated across all federally supported hospitalizations the total is fairly important. Again, an experimental trial of precertification might have helped. A trial (carried out in Hawaii) of review of ambulatory care for appropriateness of treatment indicated that such review is probably not cost effective, i.e., it costs more to conduct the review than is saved by reducing inappropriate treatment costs (*The Hawaii EMCRO,* 1973).

The treatment of patients in medical emergencies provides other examples, especially pertinent in this context, of inadequately evaluated treatments, some of which were taken for granted with some unfortunate results. Standard treatment for burns, as an instance, for many years called for administration of intravenous calcium along with massive blood transfusions, a practice now regarded as harmful because the large amounts of calcium may induce cardiac systole. The Trendlenburg position (head down) for shock victims was recommended after World War I on the basis of experience with pelvic surgical patients, and on that basis alone it was accepted as good practice for 50 years or more. It is now known that that position is wrong, the preferred position being with the patient's torso flat and the legs somewhat elevated. The Trendlenburg position example does illustrate the problem that arises when a treatment is better than some known alternatives, e.g., it is better than having the patient flat or with head elevated, but not the best alternative available. A partially effective treatment or other intervention may inhibit research to a very powerful degree.

Evaluation: begin at the beginning.

If good evaluation is to be accomplished, we believe firmly that it must be planned for, and in fact it should be *built in* during the initial stages of

program planning and development. Once they are underway, programs have a strong tendency to develop their own internal logic and momentum so that it is very difficult to probe into them to determine their effectiveness, let alone to change them. The very examination of a program from the standpoint of its outcomes becomes quite threatening. People become identified with programs and develop a proprietary interest in them at the very least. In some instances the interest becomes material. As an example of the former, it is very clear that any proposal to evaluate the performance and effectiveness of volunteer rescue squads would be likely to meet with great resistance from the squads to be evaluated. But the resistance would not be any less if one were to propose a comparative evaluation of emergency rooms operated under hospital control and those operated by contract with an outside firm of emergency physicians. The Experimental Medical Care Review Organization (Evaluation of Hawaii EMCRO, 1974) in Hawaii engendered great hostility in the local medical community when it published a study interpretable as indicating that subscribers to the Kaiser Permanente prepaid health plan might be receiving better medical care than those citizens seeking attention from private practitioners. The best way to maximize the chance that an evaluation can be properly and correctly carried out is to build it into the program plans from the beginning.

20

Evaluation is often expensive.

The potential expense of research cannot be glossed over. Program evaluation is rarely cheap, or at least rarely both cheap and good. However, one's perspective on the cost of research has to include the cost of the program or the treatment to be implemented, in some cases the cost accumulated over a good many years. The perspective also has to include some estimate of the likelihood that the change or intervention planned might actually be harmful, the likelihood that whatever bad effects might result would be reversible and at what cost, and the likelihood that a program might become a model for wide implementation. Even *very* expensive research may be worthwhile under some circumstances. For example, one group was asked to develop a plan to evaluate the effectiveness of an areawide EMS for which a federal grant of about $900K had been received. After due thought to the problems involved the planning group came up with an evaluation proposal which would have cost about $1.5 million, a result which caused a great deal of amusement and even derogation in some quarters. However, there are now more than 200 regional EMS, with many millions of federal dollars being spent, and still with very little good information on which to make a judgment of what is happening.

Many similar examples can easily be found. There was a $3 million dollar proposal to evaluate the performance of seven nurse practitioner (PRIMEX) programs, each graduating only a few trainees each year. Viewed as an evaluation of the seven specific programs the research would clearly have been dreadfully expensive. On the other hand, viewed as an evaluation of prototype programs for potential nationwide implementation, the research could have been considered a real bargain. Evaluation of 911 systems is not being accomplished, in part because the cost of evaluating any one installation would seem disproportionately great in relation to the cost of the system, say in one or two counties. Yet the aggregate cost of 911 systems across the country will be staggering, and they will all be in place before anyone discovers whether it is really a good idea or not. By that time it will be too late.

Heavy expenditures for research can also be justified when risks of bad outcomes are substantial and when those outcomes might not be easily reversed. How much would it have been worth, for example, to have done a definitive evaluation of the effects of thalidomide? Utilization of various paramedical personnel would not seem to be completely without risk, and at least some of the risks that are imaginable are also substantial, and the expenditure of fairly large sums of money to evaluate performance of paramedical personnel would seem completely justifiable. Some changes or innovations need careful evaluation, preferably in a limited experiment, because they tend to be irreversible. It seems scarcely likely, for example, that it would ever be possible to get the law changed so as to permit untrained ambulance attendants to function again, even if EMTs proved not to be any better in performance. It will be difficult, perhaps impossible, for any community to abandon its 911 system once it is in place. Nearly all of the costs are incurred in start up, and by the time the system might be found to be no better than previous systems, it would be too late. A volunteer rescue squad, once replaced by hired staff, might be extraordinarily difficult to assemble again.

To reiterate, research very often costs a lot of money in absolute terms. Whether it is relatively expensive and worth doing depend on a number of other factors, including especially whether a research effort is viewed as addressed to a specific time and space limited problem or whether it is addressed to a problem better considered as extensive in time and space.

More basic research is needed

One of the distinct impediments to the kind of research which all of us would like to see done on

EMSs—and many other health programs—is that so much basic research, preparatory research, needs to be done, and there is so little impetus and enthusiasm for doing it. We would all like to know whether trauma centers save lives, whether EMT training is worthwhile, whether it would be worthwhile to reduce rescue squad response time from ten to eight minutes. But we do not know how to measure outcomes, or even whether that measurement is possible. It is disturbingly difficult even to get basic data on emergency medical services: what proportion of ambulance runs involve unconscious victims? what proportion of ambulance runs involve multiple victims? what proportion of ambulance runs involve burn victims of what degree of severity? on what proportion of runs is basic and effective assistance already being rendered at the scene? The list is virtually endless. The answers may be available, but they are certainly not readily available, and the unavailability of answers to just such simple questions is retarding research efforts. One cannot, for example, expect to evaluate EMT treatment of burn victims if there are very few burn cases handled. Nor can one evaluate very well the handling of cases for which there is little variability, as might be the case for certain types of relatively minor injuries for which the treatment would be obvious. As yet very little is known about the way rescue teams actually function, and until that knowledge is obtained, it will be difficult to advance in other areas. Unfortunately basic research, even in applied areas, is often tedious, has low immediate payoff, has very little payoff of any kind to the agencies that are the subjects of the research, and is not very glamorous. It is, unfortunately, only critical.

Generalizability of findings

There might be some confusion created by some of the above discussion because there have been repeated jumps from local to national problems, from little to big problems, etc. It is apparent that the national interest in EMS research cannot be satisfied by purely local problems and issues. Whether a new director of a department of public safety will do a better job than his predecessor is not an item of interest beyond the locality in which the problem resides. Whether in a given community rescue squads should be kept together in teams or shifted around for convenience in scheduling is not a question of much interest in Washington, D.C. Nonetheless, we do want to affirm our belief that even local agencies would do well to have evaluators available to help determine the effects of even such limited and local changes, whether the evaluators are regular staff members or consultants. We believe that it is important for local public agencies to know what they are doing and what effects they are having. However we would also like to suggest that the perspective that one takes on a problem may determine whether it is of purely local interest or whether it has more far reaching implications. The question of replacing Chief Jones with Chief Smith is not very interesting, but the question whether replacement of Chiefs makes any difference when things are not going well is at least a potentially interesting question. One investigator has been able to show that when baseball teams change managers, performance of the team generally improves (Grusky, 1963). Could the same be true of the EMS? Similarly, the question of scheduling of rescue squad workers is of more general interest if one asks whether workers consistently assigned together function more efficiently and effectively, whether they tend to develop role specialties, and other like questions.

In any case, it should be clear that the interest of federal agencies is in research that contributes to the general body of knowledge about the working of EMS and, at least in the longer run, to the development of policies to guide federal support for EMSs. No matter how praiseworthy on other grounds, a service program to benefit a local community cannot qualify as research. Still other research is of such parochial nature and so far removed from interests of federal policy that it would not be likely to engender much interest at the federal level. For example, what sort of uniform would be most suitable for EMTs in Houston might be an issue of some concern there, but the implications beyond that community would very likely be limited and probably (many would hope) beyond the policy interests of the federal government. Research will be of greatest interest when it is addressed to problems of rather broad concern, when it promises to provide new information, when that new information will be of value in understanding the basic processes of EMS functioning and when the results are likely to be translatable into policy statements and action implementations.

Problems to be resolved

We do not want to gloss over any of the problems or limitations involved in the type of research and systematic program evaluation we are proposing here. Both the problems and limitations are numerous and severe, so much so that they remind us of Winston Churchill's comment that democracy is a terrible form of government, having as virtually its only strength the fact that it is preferable to any alternative. What is the alternative to determining whether one's treatments work? Prof. Frederick Mosteller once alleged that the only alternative to experimenting with people is to fool around with people (see Gilbert, Light, & Mosteller, 1975).

One distinct limitation of program evaluation is that administrators must often make decisions in

a time frame that does not encompass the determination of effectiveness of a proposed change. We suspect that at least some of the urgency of decision making may be exaggerated, but nonetheless, if an incompetent person must be fired and replaced, there will be no time to evaluate the effects of the replacement. The reorganization of a hospital and community health system may force changes in emergency medical services which cannot be evaluated before being made. However, in our view such problems merely reinforce in the strongest way the case for doing research and evaluation whenever it *is* possible. By having available a good data base, by having access to a fund of accumulated research, by knowing the results of evaluations of programs similar to the one being considered, it should be possible to make more intelligent, informed decisions with a much higher probability of payoff. Thus, for example, the twenty-five year research program of Prof. Fred Fiedler on effectiveness of different types of leaders in different types of settings provides at least the possibility of doing better in the replacement of an executive than merely hoping that the most available candidate will be an improvement (e.g., Fiedler, 1971). Enough is known about media campaigns to inform the public about some service that one need not start from scratch in designing an information campaign about a 911 system, e.g., we know that public service TV announcements are rarely broadcast at prime times. Whatever information is available about organizations, programs, etc., has come from research which was done when it was possible. The opportunity to do a good piece of research is not a regular occurrence, and no good opportunity should be passed up.

The work of Nathan Caplan of the Institute of Social Research at the University of Michigan has shown that there are some fairly clear limitations on the utilization of research findings in policy decisions (Caplan, Morrison, & Stambaugh, 1975). One of the clearest limitations was the reluctance of policy makers to consider the use of research not done in their own settings. That is a very serious limitation if it persists, because it is obviously impossible to replicate every bit of research in every setting. In some degree there is going to have to be an effort made to educate administrators to the use of research findings and to decrease their parochialism and sense of uniqueness and their fears of being wrong on occasion. Perhaps more stress by researchers on the more generalizeable features of their work would be helpful and that suggests again the importance of the perspective in which the work is viewed. While it is true that no two cities, nor any two hospitals, nor any two rescue squads are quite alike, it is similarly true that no two cities, etc., are entirely different. One needed and promising line of research that could be carried out as easily in the EMS field as any other is the conditions under which change occurs, innovations are disseminated, and research findings are utilized.

We would like to conclude this section by reverting to the point with which we began. The making of decisions about provision of public services is a complex matter that must take economic, logistical, political, and other realities into consideration. However, we believe that the effectiveness of a proposed change, innovation, or program is an equally vital reality which must be a factor in the decision of an administrator. We would grant that for political purposes an administrator might very well adopt a program known to be ineffective or of little worth, but that decision is better made in full knowledge of the program's lack of worth, even if the administrator then runs the risk of being considered cynical. Perhaps it is better to be cynical than to be gullible and naive. An administrative body such as a city council may not want to vote funds for a program because of fear of citizen reaction to higher tax rates, but we believe that those citizens are better served if the city council fails to enact a program in full knowledge of its actual social worth. When I buy a car, its performance characteristics is not the only factor affecting my decision, but I want to know them. Ignorance is bliss only until, inevitably, its consequences catch up with you.

References

Berger, R., & Gold, M. *Experiment in a juvenile court.* Ann Arbor, Mich.: Institute for Social Research, 1976.

Caplan, N., Morrison, A., Stambough, R.J. *The use of social science knowledge in policy decisions at the national level: a report to respondents.* Ann Arbor, Mich.: Institute for Social Research, 1975.

Earle, H.H. *Police recruit training: stress vs nonstress.* Springfield, Ill.: C.C. Thomas, 1973.

Evaluation of Hawaii EMCRO. Report under contract HSM 110—73-526 to National Center for Health Services Research by A.D. Little, Inc., Cambridge, Mass., 1974.

Fiedler, F.E. Validation and extension of the contingency model of leadership effectiveness: a review of empirical findings. *Psychological Bulletin,* 1971, *76*, 128–148.

Gilbert, J.P., Light, R.H., & Mosteller, F. Assessing social innovations: an empirical base for policy. In C.A. Bennett & A.A. Lumsdaine (Eds.) *Evaluation and experiment: some critical issues in assessing social programs.* New York: Academic Press, 1975, pp. 39–193.

Grusky, O. Managerial succession and organizational effectiveness, *American Journal of Sociology,* 1963, *69*, 21–31.

Martinson, R. What works—questions and answers about prison reform, *The Public Interest*, 1974 (Spring), 22–54.

The Hawaii EMCRO: An experiment in non-punitive peer review. Project Report. Grant No. 5 R18HS 00795 SRC. National Center for Health Services Research, Rockville, Maryland, 1973.

Weingarten, V. Report of findings and recommendations of the President's Committee on Health Education. *Health Education Monographs*, 1974, 2 (Supplement 1), 11–19.

Evaluation Research: What Is It and How Is It Done?

Linda Victor Esrov
Psychology Department
Florida State University
Tallahassee, Florida

24

The term "evaluation" is currently being used in several different ways with widely different implications for how evaluations should be carried out. While there is no one definition of evaluation that can be claimed to be the correct one, there are some evaluations that are more penetrating than others. It is important to know in what sense the term is being used when evaluations are said to be desired or to have been accomplished. In this paper Linda Esrov, an evaluation research methodologist, describes the types and levels of evaluation that are in current favor.

Over the last ten years or so a confusing variety of activities have been lumped together under the heading of evaluation research or program evaluation. This diversity is so pronounced that I assume that many people, upon picking up a volume entitled "Final Program Evaluation Report," would be hard pressed to predict much of anything about what type of information is inside. Because of this diversity authors who have tried to provide a comprehensive definition of program evaluation, one that covers all of the types and levels of evaluation activities, have been forced to produce broad generalities such as the following. Program evaluation is any assessment or information that allows one to reach decisions on programs (Bernstein & Freeman, 1975). The vagueness of this definition is a testimony to the fact that being more explicit would have excluded somebody who was doing something that he/she called evaluation research. This definition does, however, make the contribution of asserting the purpose of evaluation research or program evaluation. It has a generally agreed upon, applied purpose, that is, to aid decision-making concerning programs. However, this definition leaves unspecified at least two important considerations:

(1) the level of the evaluation (i.e., what is it about the program that is being assessed or evaluated),
(2) the methodology of the evaluation (i.e., how is the assessment or evaluation to be done).

If one includes these two specifications in a definition of program evaluation, the definition no longer refers to the multitude of activities undertaken in the name of evaluation. Instead, it defines a specific type of evaluation and consequently excludes other types. For example, one might define what is generally believed to be the most scientifically defensible type of program evaluation, namely evaluation as a controlled experiment (e.g., Suchman, 1967; Campbell, 1959; Weiss, 1972; Reicken & Boruch, 1974; Bennett & Lumsdaine, 1975), as the use of the social science methodology of the controlled experiment to assess the extent to which a program is successful in bringing about the desired changes in the target population. This can be viewed as one type of evaluation. According to this definition what is being evaluated is the program's outcome or effectiveness in producing change and the method to be used is that of the controlled experiment.

As has been mentioned, however, there are numerous definitions of program evaluation in addition to this one of evaluation as a controlled experiment. It is being suggested here that one of the reasons for the diversity is that different people are talking about different types of evaluation activities when they define program evaluation. It is also proposed that two characteristics, 1) level (what is being evaluated) and 2) methodology, vary across different definitions of evaluation research, and therefore should be useful as a means to classify different types of evaluation. Accordingly, these two characteristics will be used to develop a descriptive classification scheme that will attempt to include most of the activities that are currently labelled evaluation research or program evaluation. The rationale for such a scheme is to provide descriptive information so that one is better able to differentiate among various evaluation activities and hopefully to reduce some of the confusion that is related to the term "program evaluation". In addition to the description of different types of evaluation activities, an attempt will be made to point out each type's contributions to decision-making along with its limitations. Examples of evaluations from Emergency Medical Services will be considered within this framework.

Levels of Evaluation: What Is Being Evaluated?

Of the possibilities as to what it is about a pro-

gram that is to be evaluated (i.e., assessed in order to aid program decision-making) five will be identified. These levels are:

(1) program planning or objectives
(2) program implementation or structure
(3) program operation or process
(4) program's production of desired change or outcome
(5) program impact.

When assessing level 1, program planning, one is dealing with the characterization of the social problem area including what it is that needs improvement. This also includes the definition of programmatic elements and the setting of goals and objectives.

When assessing level 2, program implementation or structure, one is dealing with the inputs of the program such as resources, equipment, manpower, facilities, etc. Often administration is included.

When assessing level 3, program operations or process, one is dealing with the performance of daily program activities; the services delivered, the practices, strategies and intervention efforts.

When assessing level 4, the program's production of desired change or outcome, one is dealing with the overall effectiveness of the program to meet its predetermined objectives. These objectives usually relate to measuring improvements or changes in the target population.

When assessing level 5, program impact, one is dealing with outcomes that extend beyond the specific individuals who are served by the program, that is, the effect of the program at the broader community level.

In viewing these five levels of evaluation or what is evaluated, it can be seen that they evolve from the immediate consideration of deciding what form the program is to take (level 1) to the intermediate concerns of producing the program and delivering its services (levels 2 and 3) to the ultimate notion of determining if the outcomes, of both individuals and community, were what was desired (levels 4 and 5). It may be that evaluation at each of these levels can profitably accumulate to produce a particularly comprehensive program evaluation. However, even if evaluation is not to be carried out at all of these levels, it will be suggested later that a number of combinations of these levels of evaluation are very compatible due to certain methodological issues.

The importance of recognizing the level of evaluation with which one is dealing should not be underemphasized. One of the two most obvious shortcomings of many evaluation projects results from the lack of recognition of what it is that is actually being evaluated. The mistake often made is to assume by demonstrating success at one level that success has also been demonstrated at another higher level of the program. This should be recognized as an unverified assumption. For example, just because a program was implemented as planned or according to certain standards, its effectiveness in producing the desired change in its target population has not been demonstrated.

Methodology of Evaluation: How Is It Done?

In order to deal with a given level of a program in an evaluative manner one must use some means of assessing worth, value, or success. It should be recognized that an evaluative assessment is always a comparative process. There can be no absolute evaluation. If a program is asserted to be effective or successful, some type of comparison or contrast has been made. This comparison may be implicit or quite explicit. For example, on an implicit basis the comparison may be that this program is as good as other programs that one has the impression are successful or that this program is much better than one's impression of many other programs. The comparison process can also be made much more explicit. As will be discussed, the use of experimental design formalizes the need for comparisons through the use of comparison groups or control groups.

The need for comparisons in order to reach valid evaluative conclusions should be emphasized. The second of the two most obvious shortcomings of many evaluation projects is that they often claim more than their methodology can show. Many studies make what Campbell and Stanley (1966) call the "error of misplaced precision". These studies attend at great length to the collection of data concerning one program but are little concerned with the comparison of what conditions would be like or what results would be produced without the program or with an alternative program. The error is often to assume that all of the details that one has measured are *causally related* to the one program. This cannot usually be demonstrated without explicit comparison unless it is completely implausible that anything other than the program itself could have produced the results. In the realm of social programs this state of certainty does not usually exist.

Of the possibilities as to how to do an evaluation, that is, what methodology is used, four methods will be identified along with comments on their limitations and assets. The four methods are:

(1) description
(2) informal evaluation or reliance upon common sense
(3) comparison with standards
(4) experimental design.

As a method, description is meant to be taken literally. It refers to the systematic characterization or description of a situation or area of interest in accurate and comprehensive manner. In a sense, description is nonevaluative and the addition of one of the other methods (informal evaluation, comparison with standards, or experimental design) applied to description produces an evalua-

tion. Description is included separately here because the collection of a descriptive data base is such an extensive part of many program evaluations.

There are many familiar uses of descriptive methodologies. For example, case studies can be purely descriptive accounts of situations, persons, or events. Surveys seek to provide descriptions and use sampling methods so as to insure that the results are representative of a certain population. Other examples of the use of description are task analyses, job descriptions, and critical incidents reports.

The method of informal evaluation is equivalent to the application of conventional wisdom or the use of one's "common sense" in order to make judgments. Informal evaluation can be characterized by its dependence upon casual observation as the source of information and implicit goals as the criterion of value or success. It is the unsystematic use of subjective judgment to determine worth and really is the embodiment of our everyday understanding of the nontechnical word evaluation.

The problem with recommending informal evaluation is the likelihood that it will be of variable quality. There is no doubt that at times informal evaluation can be extremely insightful. On the other hand, informal evaluation can also be superficial and distorted and produce invalid decisions as a result of the reliance upon unrepresentative anecdotes and unchecked impressions. The problem becomes one of how to separate accurate from faulty impressions.

Using comparisons with standards as an evaluative method does include the important consideration of making the comparison process explicit. The measurement process itself is therefore usually very objective and the standards can usually be subjected to empirical test. As will be discussed, the validity of this approach, however, depends upon what it is that is being evaluated, (i.e., the level) and the validity of the chosen standards.

The methodology of experimental design is a purposeful and explicit approach to comparative measurement. This method is particularly well-suited to determining which of two or more treatments or programs is more effective or more successful. The classical experimental design in its simplest form incorporates two important ideas: random assignment of units (such as patients, hospitals, etc.) and a control or comparison group. As Boruch (1974) has noted, this comparison often takes one of two forms: the historical comparison, which is the basis for time series designs, compares the condition of the target group after the introduction of the program with the condition prior to the introduction. A contemporary comparison, which is the "standard" control group, makes a comparison between the target group receiving the program and a control group sampled from the same population as the target group, but not receiving the program. A comparison of the differences between these two groups is taken as an estimate of the program's effects.

Therefore, in its simplest form the classical experiment is a situation where a randomly chosen half of the units under study receives the program or treatment that is being evaluated and the other half does not receive the program. These groups are then measured on the variable of interest (for example, morbidity) and a comparison is made between the outcomes for each group. As a result of the controlled comparison and randomization of units this method has the ability to show the degree to which the measured results were attained *as a result* of the program or treatment. Thus experiments attempt to establish causal relations; e.g., was the program or treatment the cause of the observed changes in morbidity. The importance of random assignment to groups should be stressed because if a comparison group is chosen by any other method either of the following two assumptions are required:

(1) the comparison group is identical to the treatment group in all other factors except for the treatment being studied,

(2) one can correct for any of the relevant differences between the control group and the treatment group.

It should be pointed out that it is often difficult, if not impossible, to meet these assumptions without randomization.

In addition to the type of true experimental design that has just been described there are also numerous other designs which fail to meet the requirements of randomization. These are known as quasi-experimental designs and require that special efforts be made to rule out plausible rival interpretations to the hypothesis that the treatment caused the observed differences.

Classification of Types of Evaluation Research or Program Evaluation: Level (What It Is That Is Being Evaluated) X Methodology.

Now that we have distinguished among five different levels of evaluation and four different methodologies and described each of these briefly we can go on to discuss the different types of program evaluation that are produced by the combinations of these levels and methodologies. Conceptually one can envision a matrix with methodologies serving as four column headings and levels of what is being evaluated serving as five row headings.

The twenty cells that are produced are what we are referring to as "types" of evaluation. Actually this matrix oversimplifies the situation quite a bit. Some of the cells probably do not exist or only rarely. Some types of evaluation are done at more than one level and include more than one

Classification Scheme For Types of Evaluation Activities

Levels: What is Being Evaluated	Methodology: How Is It Being Done			
	Description	*Informal Evaluation*	*Comparison with Standards*	*Experimental Design*
Program Planning or Objectives				
Program Implementation or Structure				
Program Operation or Process				
Program Outcome or Production of Desired Change				
Program Impact				

methodology and therefore are defined by more than one cell. And in at least one rather important instance a type of evaluation is done at a level that is included in our matrix (program impact) but with a methodology that is not included in our matrix. This is cost-benefit and cost effectiveness analysis. There are probably other omissions but despite the artificiality of this matrix it is hoped that it will serve the useful function of structuring the following discussion and examples of types of program evaluation.

Types of Program Evaluation

Evaluating program planning or objectives. As has been mentioned, level 1, program planning, concerns the social problem area including what it is that needs improvement. If a specific program has already been suggested, this level of evaluation attempts to assess whether the contemplated action is necessary or to determine whether its stated objectives are appropriate. If a particular program is not yet specified but action is under consideration, evaluation for program planning concerns the collection of information that can help lead to the specification of objectives. As a result of this process these objectives should then be related to resolving a known social problem and meeting the needs of the group to which the program is directed. It can be noted here that in order to perform higher level evaluation activities, particularly the determination of program outcome or effectiveness, it is necessary to state objectives in terms of measurable outcomes. This should be done in the planning stage so that the program will be implemented in order to best attain these goals.

Evaluating level 1, program planning, is an issue of needs-assessment and it would appear that the methodologies of description and informal evaluation are best suited to this end. Thus needs-assessment surveys or censuses can be conducted prior to the implementation of the new program. These might utilize some type of health status indicator as a descriptive index of health as it exists prior to program implementation. The method will probably not remain descriptive but will become evaluative when present health status is compared with the health level that is desired or expected. This is probably done on an informal evaluative level but the possibility exists that there are explicit standards that can be used for comparative purposes.

In Emergency Medical Services descriptive information has often been collected regarding the unmet need for ambulance services. These data concern those patients who arrive at the emergency room with conditions serious enough to justify emergency transport but who have not received such transport. If these data show that many persons (*too* many according to an informal evaluation process) are not receiving emergency transport, they are useful to judge the necessity of a program to provide more emergency vehicles, etc. and to judge the appropriateness of this program's objectives to solve this unmet need.

In Emergency Medical Services Systems collection of descriptions for the determination of system level objectives is less likely to occur because there already exist standards of a sort, the fifteen points of the Emergency Medical Services Systems Act of 1973.

In a recent project to develop a curriculum for training Emergency Medical Services administrators a needs-assessment survey could have produced useful information for guiding the development of curricular materials. It could have been of additional benefit in helping to predict the likelihood of recruiting persons for such training at both the initial site and other proposed sites.

The evaluation of program planning and objectives is not really compatible with the methodology of experimental design. Descriptive methods such as surveys are particularly good for telling one the present state of "the world" and evaluating planning is the assessment of whether the plan and objectives fit "the world".

Thus the type of evaluation that comes out of the combination of level 1, program planning and descriptive and informal evaluation can be considered *needs-assessment.* It is unlikely that any effort at program evaluation would stop at this initial level of determining need. However, it is possible that if certain survey questionnaire items asked if persons would, for example, find more ambulance services desirable, and the response was quite favorable, then the assumption might be made as to the probable worth of the new program for increased ambulance services. The lack of any information on the objective worth or effectiveness of these services, however, makes this assumption totally untenable. This type of evaluation, namely needs-assessment, is probably well-recognized as occurring at the level of program planning.

Evaluating program implementation or structure. In evaluating program implementation, one is dealing with the inputs of the program such as resources, equipment, manpower, facilities, etc. Assessment at this level is most appropriate for what can be called *compliance-control* (Alkin in Weiss, 1972). Thus through the use of description of the resources and facilities of a program it is possible to compare whether or not the program contains the elements proposed during the planning phase, or to compare whether or not the program is in compliance with certain guidelines or standards for its structure. This type of description of structure is often required for funding purposes. One of the attractions of assessing the level of program implementation is that the information to be collected at this level is concrete and often easily obtained. However, problems arise when the assumption is made that by describing inputs, one has evaluated more than the program's implementation. Gibson (1973) has pointed out that the Federal Highway Safety Act of 1966 contained what were called "performance" or "outcome" criteria in its Standard No. 11, Emergency Medical Services. As it turns out these criteria were almost exclusively concerned with inputs or program implementation, not with outcome measures or program performance. However, the assumption that was being made, as Gibson (1973, p. 427) puts it, was that "if facilities exist, they are used, and if used they make a difference". Thus it was assumed that the inputs were related to operations or processes and that these operations necessarily produced the effective outcomes of good medical care. Similarly accreditations of Universities is often made on the basis of number of books in the library, number of Ph.D.'s on the faculty, etc. and as with Emergency Medical Services, this emphasis on resources and facilities does not necessarily provide evidence on effectiveness. Effectiveness is another level of evaluation and the assumption of the relationship between inputs and outcomes must be verified.

Thus evaluation of the level of program implementation through description and possibly comparison with standards produces what we have called *compliance control.* It does not appear that experimental design is an appropriate means for assessing *compliance control.* The misleading confusion of this level with the level of program effectiveness may be based on the use of a questionable evaluation process: the conventional wisdom suggesting that good facilities and resources will result in good outcomes.

Evaluating program operations or process. In evaluating level 3, program operations or process, one is dealing with program activities; the services delivered; the practices, strategies, techniques, and intervention efforts. It is at this level that most of the activities that are labelled evaluative occur. While not degrading the importance of knowing what operations do occur in an on-going program it can be suggested that much of what is termed evaluation occurs at this level because evaluation here overlaps considerably with management and administrative activities. As part of the Emergency Medical Services Systems Act of 1973, those systems receiving federal funding are required to include a evaluative component. This is often adhered to through increasing the visibility of those (usually informal) evaluative activities that occur as part of the program's internal management. As a result of this, program evaluation often becomes characterized as a confusing mixture of management and science.

The combination of the level of program operations and the methodology of description alone or in combination with either informal evaluation or comparison with standards can be termed *descriptive monitoring.* This is an important activity. Through the use of description at the level of program operations one can characterize exactly what activities are occurring as part of the program. Operations research and systems analysis go to great lengths to descriptively characterize what actual operations occur as part of the program, and what the organizational functioning of these operations is, including a description of the relations or links to the other parts of the system. *Descriptive monitoring* provides the information necessary to determine whether the target population of the program is being reached and whether the activities that are occurring are actually those that were specified at the planning stage as being related to the program's objectives. These are important contributions and it will be suggested that even at higher levels of program evaluation this information is valuable, if not necessary, for a comprehensive evaluation plan.

The problem that occurs with *descriptive monitoring* is related to lack of recognition of the level of this evaluation. The description of services delivered is not necessarily an indicator of program effectiveness. Those who would suggest stopping evaluation at this level make the assumption that the effort expended and the efficiency of the services are ends in themselves rather than means. Certainly an efficiently run system and the delivery of services may be necessary for program effectiveness, but they may not be sufficient. The well known evaluative criteria of ambulance response time and total rescue run time in Emergency Medical Services are problematic examples of remaining at the evaluation level of program operations.

Another rationale for stopping evaluation efforts at the level of program operations is that program objectives may not have been operationally defined in terms of measurable outcomes, or the outcomes may be uncertain or difficult to measure. Thus evaluators may rely on the use of illustrative incidents, case reports, or testimonials

to provide both description and informal evaluations of effectiveness. Again, this raises the issues of confusing the level of operations with the level of outcome.

Alternatively, evaluations may use comparison with standards in order to make the leap from the measurement of operations to the assumption of effectiveness. This is a common method used for assessing the quality of medical care. In a recent study, Frazier, Lally, and Cannon (1973) evaluated the quality of care given by emergency medical technicians by comparing the activities that the technicians performed with what they called "mandated treatments". Mandated treatments were explicit process standards of what treatment should be given if a patient presented with a particular sign or symptom complex. While this study provided some important information concerning emergency medical technicians' activities, its value as an index of quality of care is dependent upon the relationship between the standards (mandated treatments) and patient outcomes.

Medical care is also often evaluated through the use of expert judgments. This can be seen to be the comparison with standards methodology if one notes that experts are assumed to have useful internal standards or implicit process criteria of what is usual or acceptable as a result of their training and experience. Again the validity of comparing program operations with standards as measures of program effectiveness is dependent upon the validity of the relationship between the end result (e.g., patient outcome) and the operation. This validity may have been tested through earlier studies as is the case with many professional standards for which data exist clearly supporting the desirability of the operations. However, many practices go on because of tradition and professional values rather than data concerning effectiveness. As Bernstein and Freeman (1975) point out this is the case for the evaluation of school health programs where the annual physical exams for children are probably inappropriate evaluative criteria.

Thus, the level of program operations can be validly assessed through the means of *descriptive monitoring*. Experimental design is probably not necessary for this purpose. A common problem, however, is to assume that one has evaluated more than the level of operations. Procedures that compare program activities with standards not for description and compliance alone but for making judgments concerning outcomes, must recognize the possibility of invalid causal links between the activities and the outcomes.

Evaluating program outcome or production of the desired changes. The level of evaluation dealing with program outcome or the production of the desired changes has been defined as dealing with the overall effectiveness of the program to meet its predetermined objectives. As was noted, these objectives usually relate to measuring improvements or changes in the target population. For example, the objectives of Emergency Medical Services Systems may be defined as the prevention of disability and suffering in persons with injury or acute illness (Willemain, 1974). Thus assessing program outcomes in Emergency Medical Services can be done in terms of the reduction of death, disability and suffering or alternatively in terms of improving health status.

The combination of level 4, program outcome measurement with description and informal evaluation, can result in the *case study*. In this type of evaluation information is collected on the target group *only after* exposure to the program. The criteria that are measured may be appropriate operationalizations of the stated objectives or this method can also be used when objectives have not been operationally defined. In either situation, the case study provides a completely inadequate assessment of the program's effectiveness or production of desired changed. There is no explicit comparison which allows one to attribute observed changes to the program itself. The only comparison is an informal, implicit comparison with one's previous experience. As has been noted the problem with any type of informal evaluation is its unknown biases.

The methods of description, informal evaluation, and comparison with standards at the level of program outcomes can also produce what can be called *performance monitoring*. This is very much a part of operations research, and systems analysis and differs from descriptive monitoring in that the actual operationalizations of the program objectives are being assessed (level 4 rather than level 3). Often specific performance objectives are developed or projections are made as to what level of performance should be achieved within a certain time period. This type of forecasting is often made on a weak empirical basis. Comparisons can also be made with past program performance or occassionally with the performance of a similar program.

Rees (in Boruch & Reicken, 1975) makes the point that the types of information systems that are developed for management and performance monitoring are usually inadequate for the accurate estimation of program outcomes. Although outcomes are often measured there is usually no information on comparison groups who do not receive the program. Without this type of comparison it is difficult to attribute effects or outcomes to the program itself. Rees also notes that even though time series data are sometimes provided, that is, measurements prior to and after program implementation, they are too short (there are two few measurement points) to be interpreted with much confidence. Rees' final criticism of *performance monitoring* as an evaluative approach to determine effectiveness of outcome, is that it is mis-

taken to believe that the simple collection of information on program participants, without use of a research design, can produce good evaluations.

Despite the problems involved, program decisions are based on *performance monitoring*. As has been mentioned, the methodology often conforms to the comparison with standards approach (program performance is compared with some relative or absolute standard of expected performance in order to determine the extent to which program objectives are being met). There is, however, no test of other causal factors having produced these results rather than the program itself. Despite this methodological limitation, *performance monitoring* is used as the data base to plan, alter, and adjust program activities in order to increase the probability of achieving program goals.

30

The defense for utilizing information that is of unknown validity is, of course, one of administrative necessity. Program managers are faced with the need to take action on the basis of incomplete information and performance monitoring is often all there is to go on. In addition, experimentation is not the answer for all questions of validity in program planning development and management. As Campbell (1974) has noted, much of "this is mainly a matter of common sense knowing; it would be cumbersome to do an experiment on all features . . . many errors of planning are visible to the naked eye". After something is implemented, one can often see that it is not acceptable or not what was expected. Campbell uses the analogy of debugging a computer program here. It could be suggested also that if a program manager sought all of the answers to validity questions he would use much of his time and resources without delivering many services.

There is a problem though, if one's orientation is to equate evaluation *solely* with a model of continuous performance monitoring for immediate feedback to make revisions and alterations of program elements. In curriculum evaluation where this type of continuous monitoring with feedback is known as formative evaluation, this process is considered as a precursor to a summative or outcome evaluation. Thus if the real question of interest concerns the level of program outcome or effectiveness, program managers should be encouraged to go beyond performance monitoring and to introduce planned variations into their projects. There are opportunities for the evaluation of the effectiveness of different strategies and different components through the use of experimental designs. In addition, program managers can begin to collect better time series data so that if true experiments prove unworkable, quasi-experiments can be attempted.

The evaluation of level 4, program outcome, through the method of experimental design is generally considered the most appropriate way to measure program effectiveness or outcome. The classical experimental design including random assignment of subjects to the treatment-condition and a control-no-treatment condition has been described earlier along with its advantages. The most important issue is that an appropriate comparison must exist so that the measured changes or outcomes can be causally linked to the program or treatment and can not be accounted for in other ways.

A number of research projects in Emergency Medical Services have utilized the combination of outcome measurement and experimental design. For example, Wortman (1975) reports on a study by Fletcher where the effectiveness of a "follow-up clerk" in an emergency room was being evaluated. This study included measurements at both the operations (process) level and at the outcome level. The methodology was the classic experimental design. Patients who came to the emergency room were randomly assigned to either a "follow-up clerk" who phoned to remind them to keep appointments or to the usual procedure of receiving only an appointment slip. At the level of operations the clerk was successful in encouraging more people to return for treatment as compared with the control condition. And records showed that the "encouraged" patients received significantly more diagnostic tests than their control counterparts. However, when outcome criteria of health were measured, there was no difference between the two groups. This study thus suggested that there was not a causal link between health care and increased health in this situation.

A study is being conducted in Chicago by Sherman (1976) to evaluate the effectiveness of mobile intensive care units (MICUs) in reducing deaths due to myocardial infarction. This study at the outcome level is utilizing the research design of a multiple time series. This design involves a historical comparison process. A number of Chicago area communities have recently implemented MICUs and Sherman plans to gather mortality data both prior to the implementation of these units and subsequent to it to determine if the introduction of MICUs changes the pattern of these data.

One point that should be made concerning experimental designs at the level of program outcomes is that such studies are often greatly enhanced by the collection of evaluative data at the level of program operations or processes. It may appear obvious but it is a good idea to know exactly what took place during a program otherwise one may be dealing with the outcome or effectiveness of a treatment that is very different from what one thought one was examining. To illustrate this point, Hyman and Wright (1967) relate a story about the evaluation of a propaganda campaign based on the distribution of fliers. Due to a severe shortage of volunteers, however, it was never possible to distribute these fliers. Thus had the evalu-

ation taken place, the conclusion that the distribution of literature was not effective in producing the desired outcome, attitude change, would have been quite misleading. While evaluating a literally nonexistent treatment may not be too much of a threat to Emergency Medical Services, the collection of process data can provide other useful information. The following are some important uses of process information:

(1) Process information can provide data concerning unanticipated or undesirable as well as desirable outcomes.
(2) Process data can provide an independent cross-validation of the outcome effects.
(3) Process data can provide important information for estimating the plausibility of rival threats to interpretation in quasi-experimental designs.
(4) Process data can provide information for new hypotheses.

Evaluating program impact. Program impact was defined not as the equivalent of program outcome, as the term is sometimes used, but instead as the effect of the program on the broader community, those outside of the population consisting of the consumers of the program's services. Therefore what it is that is being evaluated is community outcomes.

This level of evaluation can be combined with any of the methodologies but it is most likely to be assessed through description. Thus a descriptive base that is broader than the population served by a program can be part of program impact evaluation. As Atkisson et al., (1974) point out the "social ecology" of the whole community has become an important area of concern for evaluation.

Community impact can also be assessed in a research design which is testing the hypothesis: would this community be any different if the program did not exist or if the program had taken a different form? This type of evaluation can be particularly useful if the program is predicted to produce effects at the community level. It would seem in Emergency Medical Services that a study designed to evaluate the effectiveness of categorization or health planning councils should attempt to assess community impact. Thus the effects of interest would be system effects rather than individual effects.

If it were determined that the role of Emergency Medical Services Systems appears to be to change the site of death from in the field to in the emergency room (as has been hypothesized by Gibson), a legitimate question concerns the impact on the community of these services.

It can also be suggested that when cost/benefit and cost/effectiveness analyses are applied to programs what it is that is being evaluated is program impact.

Cost/benefit analysis can be viewed as a step above the level of program outcomes both because it utilizes information on outcomes in order to quantify benefits and because it deals with social evaluations not individual evaluations. Cost/benefit analysis is an approach which attempts to quantify both the costs and benefits of programs in order to determine whether the benefits achieved by a program exceed the costs. This approach appears to be best suited to comparisons among alternatives. Since few programs can be justified at any cost, this type of analysis produces information that is relevant at the community level.

In summary, a classification scheme has been suggested which describes types of program evaluation activities in terms of what it is that is being evaluated (level) and how it is done (methodology). The five levels of evaluation considered were: (1) program planning or objectives; (2) program implementation or structure; (3) program operations or process; (4) program outcome or ability to produce change; and (5) program impact. The methodologies were (1) description; (2) informal evaluation, (3) comparison with standards, and (4) experimental design. Two persistent problems in the evaluation area appear to be lack of the recognition of the level of the evaluation and lack of recognition of the limitations of certain methodologies. Examples from Emergency Medical Services were presented and the suggestion was made that comprehensive evaluation strategies should include more than one type of evaluation.

References

Attkisson, C.C., McIntyre, M.H., Hargreaves, W.A., Harris, M.R., & Ochberg, F.M. A working model for mental health program evaluation. *American Journal of Orthopsychiatry*, 1974, 44, 741–753.

Bennett, C.A. & Lumsdaine, A.A. *Evaluation and experiment.* New York: Academic Press, 1975.

Bernstein, I.N. & Freeman, H.E. *Academic and entrepreneurial research.* New York: Russell Sage, 1975.

Boruch, R.F. On approximation to true experiments. Paper presented at Loyola Institute on Evaluation Methodology, Loyola University, Chicago, 1974.

Boruch, R.F. & Riecken, H. *Experimental testing of public policy.* Boulder, Colorado: Westview Press, 1975.

Campbell, D.T. Reforms as experiments. *American Psychologist*, 1969, *24*, 409–429.

Campbell, D.T. Qualitative knowing in action research. Address presented to American Psychological Association meeting, New Orleans, September 1, 1974.

Frazier, W.H., Lally, P.P., & Cannon, J.F. *EMT performance evaluation: A clinical trial.* Yale-New Haven Hospital, 1973.

Gibson, G. Evaluative criteria for emergency ambulance systems. *Social Science and Medicine*, 1973, 7, 425–454.

Hyman, H.H. & Wright, C.R. Evaluating social action programs. In P.F. Lazarsfeld, et. al. (Eds.), *The uses of Sociology*. New York: Basic Books, pp. 741–782.

Riecken, H.W. & Boruch, R.F. *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press, 1974.

Sherman, M.A. An evaluation of mobile intensive care units. Manuscript, Northwestern University, 1976.

Suchman, E.A. *Evaluative research*. New York: Russell Sage, 1967.

Weiss, C.H. *Evaluation research: Methods of assessing program effectiveness*. Englewood Cliffs, New Jersey: Prentice Hall, 1972 (a).

Weiss, C.H. *Evaluating action programs: Readings in social action and education*. Boston: Allyn and Bacon, 1972 (b).

Willemain, T.R. The status of performance measures for emergency medical services. MIT operations research center technical report no. 06–74, 1974.

Wortman, P.M. Evaluation research: A psychological perspective. Manuscript, Northwestern University, 1974.

Experimental Design and Causal Inference

Lee Sechrest
Professor of Psychology
Florida State University
Tallahassee, Florida

In designing research on the effectiveness of some program or other intervention the problem is to design the research in such a way as to produce data which are as unambiguously interpretable as possible. The interpretation which is desired is that a particular program or treatment definitely did or definitely did not have an effect on the outcome variables measured. In the following paper Sechrest, an evaluation research methodologist, discusses the problems that are involved in designing research which will produce convincing results.

The aim of every evaluation project should be to produce an unambiguous inference concerning the worth of the intervention being evaluated. To produce such an inference is rarely a straightforward matter, and it often involves technological and methodological issues of truly formidable complexity. However, to the degree that the final inference of worth is in doubt or is otherwise ambiguous, the purpose of the evaluation is vitiated. It is the thesis of this paper that methodologically sound experimentation is the surest way of reaching causal inferences of reasonable certainty.

An experimental study of a social intervention is devised to yield information permitting the inference of a causal link between the intervention being studied and the outcome. In the discussion which follows the experimental methods that may be employed in program evaluation are presented. While a strong case can be made for carrying out true experiments, to be defined later, in evaluating social programs, it is evident that such experiments cannot always be accomplished, and some approximations are required and may be reasonably tolerable. In the discussion which follows some of the methodological problems which, if not peculiar to program evaluation, often plague it are discussed also.

Scientific Methods of Investigation

Experimentation is not the only method of science. Cochran (1955), one of the foremost figures in development of experimental designs and their associated statistics, describe three approaches to scientific investigation: *chance observations, planned observations,* and *experiments.* Scientific inferences have often come from some very unusual happenings noted by an alert scientist. The apple falling on Newton's head, the unusual contamination of some plates in Alexander Fleming's laboratory, and the identification of vinyl chloride as a carcinogen because of the common home location of several cancer victims, are only three instances of a multitude of serendipitous observations. In contrast to chance observations are investigations which use highly detailed, planned observation schedules such as used in the Peterson study of physician performance (1956), which involved the use of highly detailed protocols for observation and the use of highly trained observers.

While scientific inference can be a product of observation, intuition and judgment are not often the basis for very firm inferences about causes and effects. Strong causal inferences are most often derived from specially contrived experiments. The word *experiment* connotes an interference with the ordinary occurrences of nature. Here we deliberately apply certain chosen procedures for the purpose of measuring the effects of these procedures. An experiment is the surest way of elucidating relationships that we are interested in observing or demonstrating. With the observational method, inferences of causal linkages derived from correlations would be hazardous and uncertain. For example, a recent newspaper story indicated that podiatrists have found that cardiac disease victims have an unusually high incidence of bunions! However, just what links bunions to cardiac disease is open to question; the podiatrists think bunion sufferers get less exercise. It would be even more hazardous if one relied upon intuition to infer causation. The precepts of science demand observable phenomena as evidence for any assertions.

Essentially the problem in evaluation research as in other areas of science is to make observations in such a way as to permit the drawing of inferences of a causal nature linking some treatment, independent variable, with an outcome, or dependent variable. Ideally we would like to be able to make an unambiguous inference, such as:

—If two hospitals of medium size are merged,

costs per unit of service delivered will go down.

—If unemployment in a given area goes up, there will be an adverse effect on average health status of residents within one year.

—If food service workers are provided with an incentive to reduce waste, there will be a decrease in waste greater than the cost of the incentive.

Unfortunately the inferences we are permitted are rarely so straightforward. More often they will be of the form:

—The merger of hospitals of medium size is often associated with a decrease in costs per unit of service delivered. (But it may have been because hospitals tend to merge when costs are abnormally, but temporarily, high.)

—When unemployment is a given area goes up, there is likely to be a decrease in average health status of residents within one year. (But maybe because the healthier people leave the community.)

—An incentive program to reduce waste was introduced into a food service, and waste went down. (But maybe because there was a change in food processing procedures during the study or maybe merely because the incentive program drew attention to the problem.)

Plausible rival hypotheses

The reason we very often cannot arrive at clearcut inferences of a causal nature is that our observations or investigations were conducted in such a way as to leave tenable or possible one or more rival explanations to the one we favor. Such rival explanations have been called "plausible rival hypotheses" by Campbell and Stanley (1963). We are all familiar with the tenets of experimentation and use them regularly in our daily life. We cannot start our car. We hypothesize that our battery is dead, and we try the lights, horn, or radio and find plenty of power. Our little experiment weakened, or even left unacceptable, the hypothesis that our battery was dead. So we go on to another hypothesis. Or a neighbor says, "I really found some good tomato plants this year. Look at them; they are twice as large as the ones I planted last year!" It is possible that he put more fertilizer on them? Have we had better weather this year? Each of those ideas is a plausible rival hypothesis to the one that the plants are superior. In the process of planning research we will be assisted greatly if we ask ourselves what alternative explanations for our findings will still be possible after we have completed our study, and we will be better able to interpret research findings if we ask what alternative explanations might account for findings available to us.

Our aim in research is to rule out as many rival hypotheses as possible as surely as possible. The problem with many types of research, and with all poorly done research, is that plausible explanations are left open and reasonable. Under most circumstances, correlational studies, i.e., studies involving natural observations, do not permit one to rule out the possibility that some underlying or third factor may account for the findings. Smokers have a high rate of lung cancer, but many people still believe that there might be some underlying factor that causes people both to want to smoke and to be susceptible to lung cancer. Some parts of the U.S. have unusually high or low rates of certain types of cancer, and maybe it is because of the mineral content of water and foods in those areas. But maybe also the areas differ in the genetic stock of residents of them, maybe people who like the particular climates or living conditions in those areas have dispositions to particular forms of cancer, or maybe some other mysterious force is operating. How could we get definitive answers? We could not, in fact, but if it were feasible and acceptable in a free society, we could take a sample of teen-age boys and teach some of them to smoke tobacco and prevent others from doing so. If we chose randomly which boys were to go in which group, in twenty years or so we would begin to find out the real answer to the smoking-lung cancer question. In the other case, we could assemble sizable groups of people and then pick randomly from them some to be sent to live in Nebraska, some in New Mexico, and in Georgia, etc. Again, in twenty years or so we would begin to get the data which would answer our question about geography and cancer. Clearly not all questions can be answered by such experimentation. It is part of the art and science of research design to conceive ways of gathering data on problems in such a way as to zero in on the right answer, even if a really high degree of certainty can never be achieved.

The problems in program evaluation are not different in kind from those posed above; the differences lie mainly in complexity and scope. Still, the aim of program evaluation ultimately is to be able to say with a high degree of certainty that whatever outcomes (or impact) are achieved, they are the result of the program itself and no other factor. We want to be able to say that it was the program itself and its particular characteristics that led to change or differences and that the change would not have occurred anyway, that differences are not attributable to the way the subjects for the study were selected for different treatments, that the results could not have been attributable to events happening outside the context of the study being conducted, and so on. In the discussion that follows, we will discuss some of the types of study designs that might be employed in evaluating programs and what the advantages and disadvantages of each are likely to be. A much

fuller treatment of this topic may be found in the now classic monograph by Campbell and Stanley (1963), virtually a must reading for any serious student of research design. A recent updating of that monograph by Cook and Campbell (1976) will also be very helpful.

The Why of Experimentation

Why do we do experiments in the first place? Well, presumably because we are uncertain about the effect of some treatment or intervention and want to make some observations that will lead to a definitive conclusion. An experiment is a way of putting a question to nature or reality. But there are other ways of avoiding or reducing uncertainty than through experimenting. At least one possibility does not even involve making any observations—logic, or reasoning. We may not be uncertain in the first place because all reason tells us is that some treatment or some course of action is good. One wag, for example, pointed to a sure cure for the problem of poverty. His reasoning was impeccable. Poor people suffer from a lack of money; ergo, give them some money, and they will not be poor any longer. The problem with reasoning is that it is so often wrong. One little error in a premise can lead to utterly wrong conclusions. A great many medical treatments that are perfectly logical are also perfectly wrong. The same can surely be said for a great many social interventions. Still, when all else fails, when there is no possibility of doing any kind of empirical study of a problem, reasoning is the reasonable thing to do.

Many interventions having to do with reduction of costs of operations may be examined in a logical manner. It requires no large scale experiment to decide that if two people are employed on a task that keeps either of them busy only a third of the time, money can be saved by eliminating one position. Still, we should be slow to jump even to financial conclusions, because very often we do not have all the information we need and do not even know that it is needed. A good example is provided by the use of one-officer police patrol cars in place of two-officer cars. It only seems logical that one-officer cars would save money since most of what police officers do, e.g., writing traffic tickets, taking non-injury accident reports, clearly does not require two officers. But a one-officer car deployment strategy doubles the number of cars needed if the same number of officers is to be available on the streets. Moreover, there are many types of calls, e.g., disturbance calls, accidents that require redirecting traffic, etc., that require two officers so that *two cars* have to be dispatched. Some police officials maintain that two-officer cars are less likely to be involved in accidents than one-officer cars; other officials maintain the opposite. The matter has not been resolvable by logic, and it is clearly going to require a fairly major research effort even to come close to a definitive conclusion.[2]

A second way of reducing uncertainty that does not require time-consuming and expensive data collection is to capitalize on the experience, preferably based on research, of others. Protocaval shunt surgery does not have to be tested in every hospital. Employment of nurse practitioners does not have to be tested in every pediatric clinic. Where data, good data, are available, they can be used as a basis for decision-making, and they should be. To do so, however, requires knowledge of the existence of the data, and some degree of expertise in interpreting the data. One or more of those factors may be lacking for any given problem or in any given setting. Where the requisites are met, though, the need for new data collection is obviated. A change in practice can be instituted and all that needs to be done is to determine whether the change seems to produce the expected results.

A third way of developing a basis for decision-making that exists in some few instances is through simulation, usually with the aid of a computer, of the projected change. For example, one group did a detailed and extensive task analysis, rather like a time and motion study, of emergency room operations, of case loads, waiting times, personnel availability and so forth. They were then able to simulate on a computer the effects of various changes in emergency room staffing such as cutting back on physicians and increasing nurses, etc. The problems with computer simulation begin with the need for a great deal of initial data collection as input for the simulation and end with the need for a considerable leap of faith in deciding to implement a change because the computer says that it ought to work. A computer can only do what it was programmed to do by some human, and how it behaves is dependent upon what was originally programmed for its behaviors. A computer may not be able to tell, for instance, that two people working together will produce less work than expected because they will spend a certain amount of time in gossip or other interpersonal affairs.

Note that even if changes are introduced on the basis of one of the factors just mentioned, there is still a need, or should be a need, to determine whether they are effective in the new setting in which they take place. The administrator, it seems to us, has only two choices once the decision has been made to introduce a change in practice or procedure: 1) the change can be *assumed* to be effective, or 2) data can be collected by which effectiveness chan be judged. We have come full circle. The need for data collection cannot be avoided unless one wants to operate on the basis of optimistic ignorance. If a decision to obtain data is made, the only question that remains is the adequacy of the data for the purpose of making a judgment of

effectiveness. That is what evaluation research methodology is all about, and that is why we experiment.

Rendering Hypotheses Implausible

Strictly speaking, we can never *prove* that a hypothesis or an explanation is the correct one. There is always some alternative that might be dredged up. What we can do is make observations that will make the most likely alternatives implausible or untenable. Under ideal circumstances all the really plausible alternative explanations but one can be eliminated, and a rather strong inference about the effect of some change can be made. How to eliminate or seriously weaken those alternatives is what experimental design is about. It is often helpful in understanding the problems that are involved to begin with some obvious, but faulty, types of "designs" in order to illustrate in a fairly dramatic way what the problems are.

Let us first note, however, the most ubiquitous plausible rival hypothesis of all: chance. No matter how well an experiment is conducted, we can never be absolutely *certain* that the observed results could not have happened by chance. If we saw someone flip a coin ten times and get heads every time, we might well be suspicious of either the coin or the way it was being flipped. But if there were a thousand people flipping coins ten times, there is a high probability that at least one of them would get ten heads in a row. There is, fortunately, through application of appropriate statistical procedures a way of telling in most instances whether an obtained finding could have occurred by chance or not. In effect, what we get is a statement of the plausibility of chance as an explanation in the form of a probability statement. Thus a statement that a difference between two comparison groups is "statistically significant" at the .01 level means that chance as an explanation of the difference is implausible since there is only one chance in 100 that a difference of the size obtained could have happened by chance. Note, however, that no matter how significant a statistical finding may be, there is always *some* possibility that the result might have occurred by chance. As a rival hypothesis chance can never be completely ruled out; it can only be seriously weakened.

Suppose a county health department, concerned with increasing rates of venereal disease, develops a special counselling program for all repeat victims and applies for funds to implement the program. A funding agency, whether a county health board or a state health department, might well ask, "Does the program do any good?" The smart administrator would have anticipated that question. There are several things the administrator might have done to prepare to answer such a question. At the very simplest level, he might have tried the counselling program on a group of VD repeaters and noted the number who returned for treatment within the following year. Let us suppose that 28% returned. What would such a result show? Unfortunately almost nothing other than that the counselling program can be operated. If we were on a board expected to produce funds for health programs, we would be inclined to ask such questions as: How many would have returned without the counselling? Data of that sort simply cannot constitute evidence for effectiveness of any program. They fall into the category of "I feed my dog these Pamby biscuits, see how healthy he is!" In their discussion of research designs Campbell and Stanley (1963) refer to the foregoing type of "evidence" as the "one-shot case study." In their presentation of different types of research designs they employ a useful notation which designates a treatment or intervention, in this case counselling, as X and a measurement or observation as O. Thus, the one-shot case study is diagramed as X O, a treatment followed by a measurement.

A slight improvement on the case study would be effected if the administrator had examined his records to determine that prior to the counselling program 40% of VD repeaters returned for treatment within one year, resulting in a one-group pretest-posttest design, diagramed O X O. However, we skeptics on the funding board might still ask such questions as:

—Is it possible that VD rates are going down anyway?

—Have operations of the clinic changed in any way that might make repeaters less likely to come in?

—Since the repeaters are clearly growing older and VD rates tend to be lower in older age groups, is it not possible that this repeater group would be less likely to contract new cases?

—Could there have been a public education campaign or perhaps a TV series dramatizing the dangers of VD during the same time period as the counselling and hence possibly accounting for the drop?

—Was the counselling program started because it was noticed that there were a great many repeaters at that time? If so, it is not likely that subsequently the number would go down anyway as these things usually even themselves out?

Each of the above questions is based on an implicit plausible rival hypothesis that might account for the findings equally as well as the counselling program.

If the administrator were able to state that, in a group of VD repeaters seen in the clinic but unable for one reason or other to participate in the counselling program, the repeat rate was about 40%, that would be termed a static group compari-

son and diagramed $\frac{XO}{O}$ the dotted line indicating that the groups were not to be considered strictly comparable as they might be if they had been selected randomly either to receive or not receive the counselling. Such a study *might* indicate that clinic procedures, community education campaigns, or whatever could not account for the findings, but that would require the assumption that the groups were really comparable to begin with. If, for example, the comparison group consisted mostly of hard core repeaters who refused to participate in counselling, then it is conceivable that their rate would be higher anyway. Such a comparison group would add very little certainty to the interpretation of the findings.

What is needed here is a true experiment in which, from a large group of eligible VD repeaters, some are chosen randomly for the counselling program while others are accorded only the usual clinic services. There are two types of experimental designs with slightly different advantages. In the pretest-posttest control group design, diagramed R* O X O, each group is measured prior
R O O
to treatment, one group is given the treatment, and then there is another measure taken subsequent to treatment. One might, for example, determine VD rates for the year prior to counselling and the year following counselling for both a treated and an untreated group. If the experimental and control groups are chosen randomly and if they are reasonably large groups, they should be very comparable at the time of the pretest. If the treatment has an effect, they should be different at the time of the posttests.

The fact that the two groups can be expected to differ only at the posttest provides a clue to the nature of the other true experimental design, the posttest only control group design, which is diagramed R X O. If subjects are assigned randomly
R O
to groups and if the groups are of reasonable size, the groups should be quite comparable on the pretest measure and there is, then, no reason to give it. There are at least two reasons for not using a pretest if one is not necessary. First, every measure costs something, and taking needless measures is wasteful of project resources. Second, it is at least possible that an experimental treatment may work differently depending on whether there has been a pretest or not, with the consequence that results of an experiment employing a pretest may be generalizable only to other settings in which pretests are used. For example, if one were interested in the effects of a lecture on subjects' knowledge about certain aspects of respiration, it is at least possible that pretested subjects would be more alert to critical elements in the presentation and that they would gain more than would be the case under ordinary conditions of conducting the course, i.e., without a pretest.

The essence of experimentation is to define experimental and control groups in such a way that they differ only in the treatment to which they are exposed. Under such circumstances if experimental and control groups differ following treatment, it can be inferred with considerable confidence that that difference was produced by the treatment. Why, then, if experiments permit such definite inferences are not more experiments done? Why is any other design ever used? One important reason is that many variables cannot be experimentally controlled, either for practical or for ethical-moral reasons. In order to be assured that experimental and control groups differ in no way other than the treatment, the experimenter has to be able to produce the treatment when he wishes or predict its occurrence well enough to be able to expose subjects to it as desired. One cannot, for example, cause the President of the United States to make a speech, but one can expose subjects differentially to the speech when it occurs. However, one cannot produce natural disasters nor even predict them well enough to be able to expose a randomly chosen set of subjects to a disaster, even if one wished to do so. The latter point reminds that some experiments would be unethical or immoral. We cannot deliberately expose subjects to risks to life and limb, we cannot abuse them psychologically for the sake of science. The long-term effects of child abuse, for example, cannot be studied experimentally; we will always be dependent upon observational data and quasi-experimental designs.

A second reason why experiments are not more often done is that preconceptions about the efficacy of a treatment often limit willingness to distribute the treatment randomly, administering it to some and withholding it from others. Although the history of medicine, along with that of most other ameliorative professions, is replete with instances of treatments once thought mandatory but since abandoned as worthless or even harmful, e.g., bloodletting, purging, it is still very often the case that a new treatment is developed and applied to a few cases with apparently great success so that any subsequent suggestions of the need for an experimental test meet immediately with the objection that it would be unethical to withhold the treatment from anyone for experimental purposes. Although Gilbert, Light, and Mosteller (1975) conclude from a review of experimental tests of medical innovations that on the whole one would be better off to have been in the control groups, convictions about the worth of new treatments develop rapidly and become quite strong. The same can be said for many treatments having to do with the delivery of health services. Mobile coronary care units, outreach programs, com-

* The Rs here are used to signify that subjects are assigned randomly to treatment and control conditions.

prehensive health care, and the like are services which are likely to be *assumed* to be valuable and hence not researchable by experimental methods. Consequently their real worth often remains unknown although great amounts of money are being spent in implementing them on a widespread basis.

There are many other reasons why experiments do not more often get done, including the fact that the desirability of and need for a well-controlled experiment is often unrecognized; but it should also be noted that a good many more experiments get planned than ever are brought to a successful conclusion. Experiments in the social arena, in real life, are not easy to do, and many a

good, well-planned experiment falls victim to various methodological and procedural ills during its course and ends up less adequate than was ever intended or even imagined. Despite the best laid plans, random assignment breaks down, e.g., because the total number of cases available is not large enough or because some higher authority insists on subverting randomization for political or personal reasons. Control groups often get contaminated when some aspects of the treatment program get implemented in the control group as well. Sometimes out of sheer carelessness subjects are transferred back and forth between groups or important changes are made in the experimental treatment in midcourse. Social experimentation is never easy, which is all the more reason to plan and strive for the *best* experiments possible. Methodological compromises in research are always in a downward direction.

Quasi-experiments

Despite the positive plea which can be made for true experiments, it is still the case that compromises do often have to be made. True experiments cannot always be planned for, and even when they are, events often force compromises that weaken them and that later demand some sort of shoring up. When, for whatever reason, it proves impossible to do a true experiment, there still are alternatives that are better than no systematic investigation at all. The so-called *quasi*-experiments are nearly always less conclusive than a true experiment because they do not permit the ruling out of *all* plausible rival hypotheses, but by careful planning of them and judicious use of information obtained, often by combining results from several studies, it has often been possible to arrive at findings which are reasonably persuasive to people willing to be persuaded at all.

However, in our view, a quasi-experimental approach to a problem usually proves to be time consuming, expensive, uncertain, and ultimately at least a bit disappointing. A good case in point is the attempt that has been made over the past twenty years to link cigarette smoking to cancer and other health problems. A long period of time has been required to reach our present position, and the expenditure of money on various investigations has been enormous. And still we are in a position of uncertainty of at least great enough proportions that those people who do not want to believe that tobacco is hazardous to health can argue with the evidence. A true experiment could never have been done, i.e., assigning on a random basis some group of youth to be taught to smoke and some other youth to be an abstinence condition, but the forced reliance on weaker alternatives and the consequences of that reliance indicate clearly the disadvantages of the quasi-experimental approach.

The point also should be made that weak or bad research is expensive at almost any price because it does not lead to any conclusions. A good case in point is the series of attempts which have been made over the years to evaluate federal manpower programs, e.g., Job Corps. There have been 24 evaluations conducted over a period during which $12.5 billion has been spent on manpower programs, and in a review of those 24 evaluations The Urban Institute concluded that the various studies which have been done are so faulty in design and execution that neither singly nor in aggregate do they provide any basis at all on which a policy maker might arrive at a decision about the worth of manpower programs (Nay, et al., 1973). That is *expensive* research. Unfortunately many, many more examples could be adduced. Whenever one can be done, one good experiment is likely to be worth more than almost any number of alternatives.

When the true experiment is not possible, there are a number of alternatives of varying characteristics and value which are very well described by Campbell and Stanley (1963). Space does not permit the explication of more than two or three examples of the designs which Campbell and Stanley present, but we would like to illustrate some of the possibilities and problems. Before proceeding perhaps it would be useful to list the most common plausible rival hypotheses which can threaten the validity of an experiment conducted without randomization, the list being taken from Campbell and Stanley (1963).

History, those events, other than the experimental variable but occuring during the same period of time, that might account for any change. For example, a television interview with a local sheriff about the 911 system could jeopardize an experimental public information campaign, especially if the program were broadcast in an "experimental" area and not in a "control" area.

Maturation, the fact that things normally change over time. There is an old saying in medicine that with proper treatment a patient will recover from a cold in about a week; otherwise it takes seven days.

Testing, the possibility that taking some measurement will in itself produce a change upon some subsequent occasion. If EMTs are anxious about performing some procedure because of its unfamiliarity, they may be less anxious and produce different results on a second testing without regard to any actual changes in skill.

Instrumentation, the changes that can occur in an instrument or recording process over time and be mistaken for experimental effects. For example, if changes are made in a record system or if criteria for eligibility for a service are changed, an unknowing investigator might be led to a mistaken conclusion. In one fire department a cutback in personnel assigned to each engine led to the upgrading of many fires from two to three alarms, i.e., more engines are dispatched in order to keep the number of *men* present at a fire at a constant level.

Statistical regression, a somewhat technical matter having to do with the fact that if cases are selected for observation on the basis of extreme scores or conditions, there is almost certain to be a shift toward less extreme values on a subsequent remeasurement. The ten "worst" hospitals in a state will almost certainly appear to have improved if looked at again in a year while the ten "best" will not look quite so good.

Selection biases, determining that some persons get a treatment and that others do not can render observations uninterpretable or misleading. For example, there is some indication that in early trials of certain surgical procedures only patients in good enough condition to survive the surgery were included in the experimental groups while the comparison groups included many patients in poor condition, thus making the surgery appear more successful than it was.

Experimental mortality, referring to differential loss of cases from experimental and comparison groups, e.g., as might occur in the comparison of a voluntary experimental insurance program with a standard program.

It is, of course, true that two or more of the above problems might exist within any one investigation and that they might interact in some ways to make the problems even worse. It should also be recognized that the threats to the validity of quasi-experiments can as easily obscure as enhance differences, thus creating the possibility that a treatment might erroneously appear worthless as well as erroneously appear valuable.

The Non-equivalent Control Group Design. One commonly encountered quasi-experimental design, and an understandably attractive one, involves comparing a group which receives an experimental treatment of some sort under conditions seen as not permitting random assignment of some subjects to a group from which the treatment is withheld. The investigator will often anticipate objections that whatever he finds might have occurred without the treatment, e.g., because of other, broader community changes. Under those conditions it is desirable to have some group with which to compare the experimental group to try to determine whether the changes found are greater than would be expected in the natural course of events. Investigators will very often cast about in search of a comparison group of some sort, usually a group with characteristics highly similar to those of the experimental group. To the extent that the groups *are* similar, then comparisons will be revealing. However, similarity must often be more assumed than demonstrated, and even where some similarity can be demonstrated, e.g., by demographic comparisons, there may be strong residual doubts if the experimental group is special in the way they were recruited into the experiment. Thus, for example, if the experimental group consists of all the employees of a factory who volunteer for a new type of health insurance program, it may be very difficult to develop any assurance that any comparison group can be formed which would be similar enough to make a conclusion possible. If, on the other hand, the experimental group consisted of the clerical workers in Division A, a comparison group formed by the clerical workers in Division B might be quite useful if there seemed to be no particular reasons why workers were in one Division or the other and if working conditions in the two Divisions seemed very much the same. The value of the non-equivalent comparison group will depend upon the case which can be made for similarity to the experimental group on factors critical to the dependent or outcome measure.

The Separate Sample Pretest-Post-test Design. Another research design that is rather frequently encountered in the health field and that illustrates some of the gains as well as shortcomings of quasi-experimental designs is the separate sample pretest-post-test design, number 12 in Campbell and Stanley's (1963) series. It very often happens that some desired intervention is difficult to apply to an isolated sample, but rather must be applied to an entire population. A good example is a public educational campaign carried out over mass media. One cannot isolate a sample to be exposed to the campaign carried out over mass media. Another example occurs if an emergency rescue service changed its dispatch procedures at some point in time, it being improbable that the procedures could be changed for only a random sample of calls. In such cases one might seek a comparison sample, e.g., a sample of individuals from a community not exposed to the educational campaign, or a sample of rescue dispatch records from another emergency rescue service. However, another possibility might be to obtain the responses from a sample of individuals in the community prior to the mass media effort and a second sample following the effort. If there is a

systematic difference between the responses of the two samples, perhaps it may be an effect of the campaign. The reasonableness of that hypothesis depends upon the confidence which one has in the assumptions that the population from which the samples were drawn did not change over time and that no other events occurred in the community which might have accounted for the response change. Thus, for example, in a survey of business firms concerning their victimization by crime, if the time elapsing between the first and second surveys is very long, the population of businesses available to be surveyed may have changed as some businessmen move out and others move in. Or if unemployment rates change from the time of the first to the second survey, crime rates may change quite independently of any police activity and either obscure or enhance the apparent effects of a police program. The separate sample pretest-post-test design is obviously not ideal, but it may have some utility when elapsed time is brief and when, luckily, there do not appear to be any dramatic intervening events which might have produced the apparent experimental effect.

Time series designs. One additional design which may be useful to note is the time series, a research design which can be implemented when one has an opportunity to make a series of *baseline* observations prior to the introduction of some programmed change and a subsequent series of comparable observations. For example, if a hospital emergency room wished to institute and test a new method of handling possible fracture cases in order to minimize unnecessary radiography, if records on radiographic procedures and positive and negative results for discovery of fractures were available by week for a period of a year prior to the change and could be accumulated weekly for a year or so following the change, there would probably be adequate data for a time series analysis of data. Any change from the pre-experimental to the experimental period might well be attributed to the intervention. However, the interpretation of findings is often not simple. To begin with the number of observations or data points needed on either side of the intervention is sizeable in most cases because of the fluctuations which normally occur and have to be dealt with. Seasonal changes or other cyclic changes pose problems, e.g., in a wintry area there might be many more cases in the winter with possible changes in base rates of genuine fractures, obvious fractures, or whatever. Moreover, if the experimental change only has a gradual change because of being phased in or because of taking time to develop, the gradual change in the post-intervention period may be difficult to interpret as an effect of the change rather than as a naturally occurring change. One would also want to be assured that *only* the critical change occurred during the intervention period. Thus, for example, if not only the method of processing fracture cases but the radiologist changed, the effects might be difficult to disentangle.

It is also apparent that an experimental intervention may have a wide variety of effects, and those effects will differ in the case with which they may be detected. For instance, the following are but some of the effects possible (see Fig. A):

(a) The initial effect is small but cumulative, as might be the case for the effects of a training program on income. The effects in the early years would be small but might well grow in size over time.

(b) The initial effect is fairly marked, but there is a fast return to original levels. An example might be the effect of a refresher training program in the schools, with personnel showing an immediate and perhaps substantial improvement in performance but followed by a quick loss and return to normal behavior.

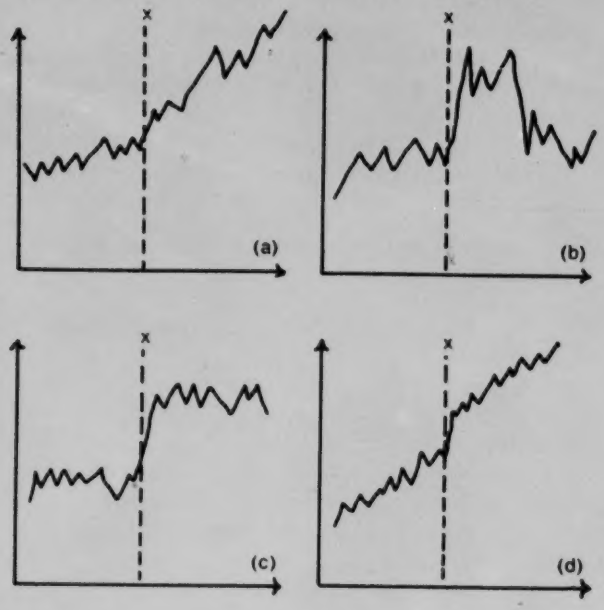
(c) There is an immediate, discrete change which is maintained over time, e.g., the instituting of an improved communications system might have an immediate effect on rescue response time with little if any further change.

(d) In a situation in which some behavior is changing gradually over time, there is an experimental intervention which slightly displaces the level of the behavior being observed without having any effect on rate of change. An example here might be the effect of some brief training program introduced in the context of gradually improving skill, such as might occur if a group of EMT students were shown a couple of nonobvious handy tricks in the handling of some items of equipment.

The above are only some of the possibilities; there are many more. Detection of changes in a time series is not an easy task, and the statistical tools needed for that detection are still in the process of being worked out (cf., Glass, Willson, & Gottman, 1975). Interpretation of a time series can often be improved if *multiple time series* can be prepared, e.g., if a comparison group is available with the experimental intervention introduced at a different point or if a comparison group never exposed to the experimental intervention can be studied. Such comparison groups can help to rule out the possibilities that factors extraneous to the experiment, such as broader community changes, mass media campaigns, maturational processes or whatever might have been responsible for the observed changes.

The requirement of rather long pre- and post-experimental observation series represents a fairly stringent limitation on the usefulness of time

Fig. A. Different Time Series Outcomes



series designs since it is not often the case that it is possible to plan for and collect data weekly for up to a year prior to and subsequent to an experimental intervention. However, there are many cases in which ongoing records may be exploited in order to obtain baseline data so that the experimental intervention can be implemented immediately, the limitation being that no change in recording procedures can have occurred or be tolerated from the beginning of the baseline period to the end of the experiment. If that requirement can be met and if the records contain information satisfactory for judging the success of the program, the time series design can be quite useful and often a reasonable substitute for a true experiment.

Accepting the null hypothesis.

It is in the nature of the evaluation of experimental treatments and interventions that one very often wishes to be able to demonstrate that the null hypothesis is tenable, i.e., that it is reasonable to believe that two treatments *do not* differ in outcome. That is particularly likely to be the case when one wishes to show that a new and simpler or less expensive program produces results as good as those produced by an established program. It is not necessary to demonstrate that the new treatment is *better* than the old one, only that it is equally as good. For example, paramedical personnel only need to be able to handle medical problems as well as more expensive physicians; a new and simpler suture need only be as good as the established procedure; a six week training program need only be as good as a ten week program. In traditional science there has been a predominant concern with mistakenly accepting a hypothesis which will later prove to have been wrong, because traditional science proceeds, and can afford to proceed, in a gradual, orderly manner, with findings being checked regularly by other investigators. However, in evaluating social programs it may be equally as harmful mistakenly to conclude that a program is ineffective as to conclude mistakenly that it is effective. Once a program is shown, however erroneously, to be ineffective, it may be abandoned and never tried again.

There are serious problems involved in attempting to show that two programs or treatments are equal in their effects. In the first place, strictly speaking it is improbable that any two treatments are exactly equal. Consequently, the likelihood of determining that they are unequal will depend on the precision with which the experiment is done and the number of cases studied. However, the more carefully an experiment is done and the larger the scope of the study, the more likely it is that a difference will be found but that the difference will be of trivial practical importance. Conversely, the smaller and more carelessly done an experiment is, the greater the probability that the conclusion that two treatments do not differ will be reached. The difficulty is that the conclusion that there *is* a difference can usually be reached with a fair degree of certainty; the conclusion that there is no difference is almost always more weakly supportable.

Still, investigators, and the consumers who use their work, do often arrive at acceptance of the likelihood that there is no practical difference between two programs or treatments. The research outcomes associated with that sort of a conclusion need to be better understood, but several factors seem to be involved in acceptance of the "no difference" conclusion. First, acceptance of the null hypothesis is facilitated by fairly large scale, carefully conducted studies. If one wished to be able to conclude that paramedical personnel can handle certain emergency procedures as well as physicians, the study should not be carried out on a small number of paramedics and physicians, nor should it be undertaken without careful attention to measurement problems, definition of cases, etc. Second, general acceptance of the null hypothesis is more likely if the conclusion of no difference has a strong, logical inferential base. It is easier to believe that two programs are equal if there is no powerful reasons to believe that they should be different. One might well believe that general surgeons would do equally as well as specialists in carrying out routine appendectomies; it would be difficult to believe that they would do as well as specialists in carrying out neurosurgery. Third, the null hypothesis is rendered more accceptable if a large number of widely varying measures showing no difference are obtained. If only one or two variables are studied, it is easy for the doubter to insist that a more assiduous search for differences

would have uncovered them. In the Kansas City police patrol experiment, for example, it was concluded that types of patrol do not differ in their effects. That conclusion is sufficient to warrant changes in patrol strategies to capitalize on opportunities for redeployment of personnel. The persuasive feature of the results is that *many different* possible outcome measures were examined, and there was no consistent pattern obvious in the few differences that were found.

It is not easy to gain acceptance of the null hypothesis, and it can never be proven, but it is not impossible to establish it as a reasonable conclusion when that seems desirable and consistent with the findings.

In favor of strong treatments.

If one has a program that one believes to be effective and if one wishes to establish that effectiveness by an experimental trial, there is one recommendation which, above all others, is likely to maximize the chances of getting the desired outcome. That recommendation is to devise and implement the treatment in a strong form. Probably as much as any other factor it is the weakness of experimental treatments that forces us to the conclusion that they are of no value. For example, it is nearly pointless to attempt to evaluate a training program that is poorly planned, carried out by inexpert instructors, and that is ill attended by trainees. It is true that those might be characteristics of eventual implementations of the program when it is actually put into practice, but ordinarily we want to know whether a training program will be effective when it is done right. After that has been established, it may then be worth determining whether inexpert instructors can carry out the training, etc.

If a treatment is delivered in a strong, optimal form, then conclusions are likely to be fairly clear cut. The program will either produce sizeable effects which will be evident in spite of design and measurement problems, usually without the need for any fancy statistics, or it will be clear that the treatment does not do very much. If it does not work well in its strongest form, it will almost certainly not do anything at all under field conditions.

Even in simple pre-experimental designs such as those involving a pretest, a treatment, and a post-test given to one group only, i.e., O x O, a striking change, especially if it is consistent across all the cases, will often be quite persuasive. If almost no trainees can do CPR properly before a training program and almost all of them can do it very well afterwards, no control group would be needed. However, if the difference is not great, i.e., the treatment does not have a strong effect, the possibility that the pre-test alone might have produced the final difference might not be unreasonable. Or if some new burn treatment seems to work well on just about all cases on which it is tried, a case for its effectiveness may be made even despite the absence of a control group. However, the problem is to develop a strong treatment and to be able to deliver it consistently. Unless one is quite confident of being able to meet those criteria, it is much better to rely on more powerful experimental designs with comparison groups.

Feasibility of Experimentation in Social Action Programs

How feasible and useful are even such quasi-experimental designs in the context of social action programs? Boruch (1974) has documented more than 200 experiments which illustrate the variety of social programs which have been subjected to experimental field test. A number of interesting approaches have been used in these experiments in order to obtain randomized assignment. Campbell (1969) argues that randomization might be very reasonable to use in the social setting. The randomization unit might be persons, families, precincts, or large administrative units. Where resources are scarce and are not available to all, randomization is perhaps the most democratic way of making them available or testing them in social programs. The necessity of introducing pilot projects and staged innovations also permits the use of random assignments as the best way of assuring equality and fairness to all social groups.

Despite all this, it is often the case that social action programs are unable to find appropriate random groups to serve as controls in experiments. In such situations, it would be appropriate, in a quasi-experimental situation, to find reasonably comparable and equal comparison groups. There are obvious problems with this, for service must be denied to certain sectors of the constituency, which results in the problem of most policy-makers wanting to assign people to treatment on the basis of their professional or political knowledge and experience. Such expediency destroys randomness or comparability and makes for difficult generalizations. Of equal importance is the problem of obtaining suitable controls and the social problem of dealing with angry, aggrieved, and distraught subjects who have been treated as controls with placebo treatments. Social action programs tend to hold out high expectations and considerable political commitments and biases due to the preconceptions and honest convictions on the part of their proposals. In such situations, administrators often find themselves trapped in advance in the need to prove the efficacy of the reform that has to be evaluated without being able to conduct an honest experiment to find out its true value (cf. Campbell, 1969). Such political pressures need to be handled with honesty and forthrightness. It would be wrong to use biased analysis in

order to demonstrate the usefulness of a reform that has been implemented.

Perhaps the most difficult fact for administrators and policy-makers to accept is that single experiments rarely prove or disprove the utility of a particular approach. The essence of good research design and statistical analysis is to be able to demonstrate that one and only one known variable could reasonably have produced the observed outcome, but any one study is likely to be so narrow or specific in the program tested, or population studied, or outcome observed that any final, unequivocal conclusions would almost always be unwarranted. That is a state of affairs that can prove very frustrating even to a program evaluator, let alone to an administrator who must make a decision. Scientists generally hope that a cumulative model might be used in social action experiments in order to demonstrate their long-term utility. The recent experience of evaluation of social action programs has demonstrated a lack of comparability of outcomes from different programs. Little seems to be done to insure that one program will take off from and utilize the experience and findings from a previous one. The very nature of large-scale investments in society requires that little overlap occur particularly where redundant and not so useful approaches have previously been tried. Thus, later programs tend to be essentially new and thereby give the impression that previous approaches have been condemned by implication. The fact is that little information tends to be gathered about previously tried approaches. Thus, the process of successive approximation is hampered.

A note about correlational studies.

There is probably no methodological and epistemological warning more often encountered than that "correlation does not equal causation." There is probably also no warning more needed. The medical field has many areas and problems that are recalcitrant to good experimental design, whether for practical or ethical reasons, and in those areas the temptation at least to collect correlational data is seemingly irresistable. Many of the correlations are fascinating enough, but few of them provide any basis on which to make policy, and not a great many more provide any basis for improved understanding of the basic processes which are at work in the field. This is not to insist that correlational data should never be collected, nor that such data are invariably worthless. Rather it is to serve as a reiteration of the warning and an encouragement to try to think through in advance the implications of a study involving correlational data.

Perhaps it is worth a line or two to explain that by correlation is meant the observation of covariation, of the relatedness of two or more variables. A correlation may involve observation of two variables as they change over time, or it may involve the values of one variable as a function of the values of another. For example, weight and blood pressure may be measured and correlated in a single individual over time, let us say by obtaining measures of both on a weekly basis. Alternatively weight and blood pressure may be measured at the same time in a number of different individuals. Correlations may be positive, meaning that a large value on one is associated with a large value on the other, with medium and small values being similarly associated. Blood pressure and weight are likely to be correlated positively in a large sample of persons. Correlations may also be negative, meaning that a large value on one is associated with a small value on the other and vice versa. Correlations between age and health status are likely to be negative, i.e., older persons have worse health. Correlations may also be essentially zero, i.e., indicating no relationship. There is probably no correlation between height and occurrence of myocardial infraction in adult males. Correlations may vary from rather large, indicating strong relationships to near zero, indicating weak relationships.

The point of the above is to indicate that correlations *only* indicate that two sets of observations are related in the sense that the values of one are some function of the values of the other. There is no indication from the correlation itself *why* the relationship exists. The assumption may or may not be correct or even reasonable. There is usually no way to be very sure without a great deal of additional information, and even then, as the smoking-lung cancer debate informs us, certainty is limited.

The problems with interpreting correlations can, perhaps, best be illustrated with some examples:

—It has been found that the more often a surgeon performs a given procedure, the better the results he gets. Should we then encourage surgeons who do not operate very often to do more surgery? Or is it possible that the better a surgeon is, the more referrals he gets?

—It has been found that teaching hospitals produce better outcomes for a wide variety of medical and surgical cases. Should we then encourage all hospitals to institute teaching programs. Bigger hospitals also get better results. Should smaller hospitals add beds?

—One study reported that the faster the travel time of a rescue squad from the scene of the emergency to the hospital, the lower the probability of survival of the patient. Should emergency vehicles then travel slower? Or

isn't it possible that the more desperate the case, the faster the driver will go?

—The Statistical Bulletin of Metropolitan Life Insurance Co. has reported that among major league baseball players third basemen have had the lowest mortality ratios, and pitchers and first basemen have had the highest mortality ratios. Is there a clue there for the parents of Little Leaguers?

The above examples were deliberately chosen as somewhat extreme, but they do illustrate the hazards of attempting to interpret correlational data. More subtle examples could as easily have been chosen, a representative one being the observation that the more years of experience a policeman has, the more cynical he is. Does police work breed cynicism, or do only the cynical survive in the police force? Experienced hang glider pilots have more fatalities than the inexperienced. They probably also fly more and take more risks. Teenage boys have more auto accidents than girls? More reckless? Less skilled? Or is it because they drive more miles?

It is true that more powerful statistical techniques for dealing with correlational data are currently being developed and studied, but their use is as yet of questionable value. Our best judgment at this time is to avoid trying to base conclusions about causal relationships on the basis of mere association between variables.

References

Campbell, D.T. & Stanley, J.C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1963.

Cochran, W.G. Research techniques in the study of human beings. Milbank Memorial Fund Quarterly, 1955, 33, 121–136.

Cook, T.D. & Campbell, D.T. The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.) *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976, pp. 223–326.

Gilbert, J.P., Light, R.J., & Mosteller, F. Assessing social innovations: an empirical base for policy. In C.A. Bennett — A.A. Lumsdaine (Eds.) *Evaluation and experiment: some critical issues in assessing social programs.* New York: Academic Press, 1975, pp. 39–193.

Glass, G.V., Willson, V.L., & Gottman, J.M. *Design and analysis of time series experiments.* Boulder, Colo.: Colorado Associated University Press, 1975.

Kelling, G.L., Pate, T., Dieckman, D., & Brown, C.E. *The Kansas City preventive patrol experiment: a technical report.* Washington, D.C.: The Police Foundation, 1976.

Nay, J.N. et al. *Benefits and costs of manpower training programs: a synthesis of previous studies.* Washington, D.C.: The Urban Institute, 1973.

Peterson, O.L. An Analytical study of North Carolina general practice, 1953–1954. *Journal of Medical Education,* 1956, *31,* 1–165.

Footnotes

1. The author is indebted to Ayres D'Costa for assistance and advice in preparing this paper.

2. A recently completed but not yet published Police Foundation study carried out in San Diego indicates very strongly that one-officer cars are safer and more efficient than two-officer cars.

Social Attitudes and Program Evaluation

Russell D. Clark, III
Associated Professor of Psychology
Florida State University
Tallahassee, Florida

There is probably no one approach to evaluating programs that is more often and more widely used than attitude measurement. Attitudes of trainees are assessed, public attitudes are tapped, attitudes of administrators are inquired after, and so on. Yet, as this paper makes clear, there are serious limitations to the usefulness of attitude measures, and evaluators should probably never rely solely on those measures.

The concept of attitude has been regarded as the most distinctive and indispensable concept in American Social Psychology (Allport, 1935). In fact, it is today the most widely used single term in all the behavioral sciences (Berkowitz, 1972). The original impetus for the study of attitudes was, and is, that they are believed to have something to do with how people act or behave. For example, the statement "the actions of the individual are governed to a large extent by his attitudes" explicitly assumes that what people say is a good indication of what they will do. In theory by the use of well-constructed questions and answers to them it is possible to obtain a great deal of information about an individual's past actions, his or her current beliefs and even intended future actions in a relatively short period of time, and then use this information to predict what the individual will in fact do in a particular situation.

Guided by these assumptions, social psychologists have gone about investigating the attitudes of a large part of the world's population. For example, attitudes about politics, race, war, money, work, sex, religion, communism, health, and so forth are constantly being reported in sources ranging from scholarly articles and books to daily newspapers. This information is not only made available to nearly everyone, but it unquestionably affects our lives in important ways. Politicians often change their views (at least as verbally expressed) to conform to the mood of the people as revealed by opinion polls. It is not even uncommon to find the latest returns in politicians' pockets. Economists study consumer buying intentions, and businesses spend millions of dollars trying to find out the public's reaction before either naming a new product or finding out what is the best way of presenting the product so that many people will actually buy it. In fact, the concern of knowing what people's attitudes are is so pervasive that it runs throughout the personal, private, and public sector of our culture. The importance of attitudes as a concept is further reflected by the fact that we often change our own attitudes in response to information about attitudes of others.

So, the social psychologist and the layman are alike in their interest in attitudes because they are thought to provide a basis for predicting overt behaviors. It is further assumed that attitudes can accurately be measured.

The organization of this paper is as follows: (1) a consideration of what is meant by attitudes; (2) a discussion of how social psychologists go about measuring attitudes; and (3) a careful look at the fundamental assumption underlying the study of attitudes.

What is meant by attitudes?

Nowhere is there more disagreement in social psychology than in the definition of an attitude. In 1939 there were 30 separate definitions in use. Today there are probably more than 100. Rather than dwell on the numerous different definitions of attitude, I will define attitude and its characteristics in a way that most social psychologists would agree with. By attitude is meant a disposition to respond to some social object in a negative, neutral, or positive manner, i.e., one is set to respond for or against something. That something may be a system of beliefs, political party, automobile, certain other persons, an institution, group, value or ideal, or one's own body. Attitudes have the following characteristics:

1. *Consistency.* The most basic and fundamental evidence for attitudes is a pattern of consistency in responses to some social object. Let us see what is implied by consistency. Suppose one day during lunch you observe a man being rather arrogantly rude to his waiter. Why? Well, perhaps the man is in a bad mood, perhaps he just lost his job or loved

one, or perhaps the waiter just delivered cold soup and a warm martini. Now suppose further that during the next ten days you eat lunch in this restaurant and everyday you notice the same rude behavior toward whoever is serving the man. You might conclude that the man feels superior to waiters. If so, what you have done is infer an attitude from consistent behaviors (rudeness) to some social object (waiter). However, if you were able to observe the same man in different settings and found that he displayed this same consistency of rudeness toward a wide variety of people, you might conclude that he feels superior to most people. That is what is meant by referring to an attitude as a pattern of consistency in responses to some social object.

2. *Acquired.* Attitudes are not innate; they are acquired or learned. Attitudes are not transmitted through the genes. Infants do not arrive in the world with preferences for a particular social, political, economical, or religious orientation; rather an individual's dispositions toward social objects is a result of the individual's prior experiences. Whether one feels positive, indifferent, or negative toward a particular social object depends upon prior experiences with that object. For example, there is a tendency to like those social objects which have led to pleasant consequences in the past and to dislike those social objects which have led to unpleasant consequences. Pleasant or unpleasant consequences may occur as a direct result of interacting with a social object or they may occur vicariously as a result of observing others verbally expressing pleasure or discomfort when engaged in interaction with an object. Attitudes can also be taught directly, e.g., as when parents teach their children to look favorably upon some system or religious beliefs. It should be apparent that an implicit assumption involved in viewing attitudes as being a function of learning is that the formation of attitudes is largely a result of the environment in which the person lives. More specifically, persons who have lived together in a particular environment will hold attitudes more similar to each other than will persons raised in different environments. Thus, on the basis of being able to identify the political climate of a nation, state, or different locales within a given state, it is possible to predict with a fair degree of accuracy whether a conservative, moderate, liberal, candidate will be elected to office. Similarly, one can predict how individuals will respond to numerous social issues.

3. *Stability.* Once formed, attitudes are stable and endure beyond the immediate time and place. Attitudes are usually thought of as relatively enduring. They are not necessarily permanent, but they are regarded as fairly stable from one day to the next or until some reason for change occurs. Examples of occasions for change would be when an individual is no longer rewarded for expressing a certain attitude, when an individual encounters new experiences which are inconsistent with prior attitudes, or when an individual is exposed to new information concerning the attitudinal object.

4. *Structure.* Attitudes have a conceptual or cognitive structure. By conceptual or cognitive structure is meant that an individual has beliefs or opinions about attitudinal objects, e.g., women are more emotional than men, examinations test only a small part of what we know, individuals on welfare are lazy, politicians tend to be dishonest, doctors care more about money than the welfare of the patient. Our beliefs and opinions tend to be consistent with our affective dispositions toward attitudinal objects. If one is favorably disposed toward a particular attitudinal object, beliefs regarding that object are likely to be positive; if one is unfavorably disposed towards the same object, beliefs tend to be negative. A person is scarcely likely, for example, to have a very positive attitude toward a certain hospital emergency room and also believe that the physicians there are incompetent. Similarly, having positive feelings toward a given object will usually lead to an expectation of consequences, whereas negative feelings toward the same object lead to expectations of negative consequences. For example, a person who is prejudiced against blacks would be more likely than other persons to believe that allowing blacks to move into white neighborhoods would lower property values, lower the quality of education, and make the atmosphere of the community less pleasant.

5. *Intensity and extremity.* Attitudes vary in intensity and extremity. Intensity refers essentially to the strength with which an attitude is experienced and extremity refers to degree of favorability or unfavorability an individual appears to have toward the attitudinal object. Attitudes vary from low to high intensity and from low to high extremity. The pattern of consistency in responses to a given social object should be greatest when the intensity and the extremity of feelings toward the object are strong. As intensity and extremity decrease a person is likely to be less consistent in his responses to the object. Most people, for example, probably have generally favorable attitudes toward emergency rescue services in their communities, but since direct experience with those services is limited, most public attitudes are probably rather poorly formed and are neither intensely held nor extreme in position. Thus, one could expect a fair amount of inconsistency in such attitudes, e.g., believing that ambulance personnel are generally competent but that they may discriminate on the basis of race or social class. Weakly held attitudes are also more susceptible to change so that a single unfavorable event involving an ambulance company might have a fairly extensive effect on community attitudes.

Ways of measuring attitudes.

Before it is possible to study the formation of attitudes or attitude change, and certainly before an individual's behavior can be predicted, it is necessary to be able accurately to measure attitudes. Not surprisingly, then, social psychologists have spent a great deal of time, effort, and money in formulating and developing measures of attitudes. The most common approaches to attitude measurement are self reports, indirect methods, physiological measures, and observational methods.

Self reports. Without question the most common way of measuring attitudes is simply to ask individuals what their attitudes are. The typical procedure involves asking individuals to complete an attitudinal questionnaire which contains numerous positive and negative statements regarding attitudinal objects. The subject is asked to agree or disagree with each item or, preferably, to indicate how much he agrees or disagrees with each item, e.g., strongly agree, agree, indifferent, disagree, strongly disagree. The underlying assumption in the latter case is that an individual who agrees is less favorably disposed toward the object than an individual who strongly agrees. Similarly, an individual who merely disagrees is presumed to be less negative toward the object than an individual who reports strong disagreement. After the questionnaire is completed the investigator merely sums the scale values and arrives at an overall index expressing favorability or unfavorability toward the attitudinal issue. Thus, based upon self reports obtained from individuals, social psychologists attempt to predict how a given person will behave when confronted with a particular social object.

In developing attitudinal questionnaires the investigator assumes or determines that the individual items are either positive or negative concerning a social object and that if individuals agree (disagree) with one particular positive item they will tend to agree (disagree) with all other positive items. In general, these assumptions are correct. Persons judging items with respect to a particular issue can agree on which items favor the issue and which do not. Moreover, research on attitudes has shown that if an individual is favorable toward one pro item, he or she tends to be favorable toward other pro items, and the converse is true for con items. In short, psychologists have been able to develop questionnaires incorporating both pro and con items on a given attitudinal issue, and there is a tendency for individuals to be consistent in their agreement or disagreement with the individual items.

In addition, the self report methods of measuring make two additional, key assumptions. First, it is assumed that a person knows how she or he feels about a particular social object. Second, the person must be assumed to respond openly and honestly to the items. These two assumptions are simple and intuitively appealing. In order to predict accurately a persons' behavior, the person must know what his attitude is and must honestly report it. To the extent that these two assumptions are not met, predictions will be poor.

Unfortunately, the validity of the last two assumptions has plagued social psychologists from the beginning. People apparently do not always know how they feel about social objects, and more importantly, even if they do know, there are many reasons why individuals either will not reveal their attitudes, or, in fact, will give deliberately misleading answers. In our culture responses to attitude questionnaires are affected by a *positivity* effect and social desirability. By positivity effect is meant a general tendency, everything else being equal, to say nice things rather than negative things about the other people. In most experiments which have been designed to affect the liking or disliking of one person for another, the liking is stronger than the disliking. Also, there is a strong tendency for individuals to give socially desirable answers. That is, when an investigator is trying to get a measure of a socially disapproved attitude, there is a strong tendency for respondents to give socially more acceptable responses. For example, in many segments of our society it is not socially acceptable to express negative attitudes toward blacks, Mexican-Americans, Italians, women, etc. Yet, many Americans clearly do have negative attitudes toward one or more of these groups, so that when confronted with a statement such as "I dislike being around blacks," "I think blacks are inferior," or "Women should stay in the home," etc., they will tend to give neutral or slightly positive responses even when in fact their attitudes are strongly negative. Here is the main problem. Whereas a social psychologist wants answers to reflect true feelings, respondents are usually concerned with what others will think of them.

The social psychologist's problem is that he seldom really knows whether the subjects' responses are genuine or a result of social desirability. Giving false responses to make themselves look good is most likely to occur when respondents know that some other person will become aware of what their attitudes are. To alleviate this problem, social psychologists tend to administer their questionnaires in large groups in which it is virtually impossible for the subjects' responses to be identified. However, even under these circumstances there is reason to believe that subjects still tend to respond on the basis of what is socially desirable. For example, one of my colleagues, Dr. J. Brigham, has been interested for the last eight years in whites attitudes toward blacks. He has had to give up several research projects because he

cannot find very many "prejudiced" individuals in Tallahassee; he cannot find subjects who will give a self report indicating "I dislike blacks." Of course, this could mean that there are no prejudiced individuals in Tallahassee, although given the hiring and residential practices of our city, we are dubious in the extreme of that proposition. A much more likely explanation is that many persons are responding more on the basis of what they know society wishes them to say than on the basis of their own true feelings.

In spite of these limitations, self reports are the most popular and frequent way of measuring attitudes, a fact that will continue to be true because compared to other approaches, self report measures are easy to develop and administer, and they are economically feasible. At the same time we must constantly keep in mind that people are not always in touch with their dispositions, and, even when they are, they will not always give completely truthful responses, particularly when they are concerned with being evaluated. We are still looking for a satisfactory solution to these problems.

Indirect methods. The indirect approach to measuring attitudes involves exposing an individual to a relatively unstructured or ambiguous stimulus situation. A person's responses to a properly chosen ambiguous stimulus are assumed to reflect his or her attitudes. For example, Haire (1950) presented the following shopping list made out by a hypothetical woman to a sample of housewives:

1½ lbs. of hamburger
2 loaves of Wonder bread
bunch of carrots
1 can Rumford's baking powder
Nescafé instant coffee
2 cans Delmonte peaches
5 lbs. potatoes

The other half of the sample were presented with the same list except that "1 lbs. Maxwell House coffee (drip grind)" was substituted for Nescafé. Each respondent was asked to look over the shopping list and then to write a brief description of the personality or character of the woman who had made out the list. The differences between the descriptions of the hypothetical woman who bought Nescafé as compared to the one who bought Maxwell House coffee were rather striking. Approximately half of the women who read the list containing the instant coffee described its buyer as lazy and failing to plan her household purchases well; the woman who bought the drip ground coffee was rarely described in these terms. In addition, the woman who purchased the instant coffee was more often seen as a spendthrift and a poor wife. Moreover, a check of the pantries of the respondents showed that most of the women who described the buyer of the instant coffee in unfavorable terms did not actually have instant coffee on their shelves, whereas those who did not describe her unfavorably were much more likely to have instant coffee. In short, it seemed that interpretation of the decision to buy instant coffee was influenced at least as much by attitudes about what constitutes good housekeeping as by reaction to the flavor of instant coffee. These attitudes might not easily have been elicited by a direct approach.

Other investigators have used sentence-completion tasks as indirect measures of attitudes. Kerr (1943) studied the national stereotypes held by the English people by presenting individuals with the following sentences to complete:

The thing I do admire America for is. . .
The trouble with America is. . .
When I think of the Russians, I think of. . .
If the British and Soviet armies fight side by side they. . .
If you invite an American to your home he may. . .

Burwen, Campbell, and Kidd (1956) employed an incomplete sentence test as one of a number of measures of attitudes toward superiors and subordinates in an Air Force population, with sentence parts such as:

He never felt comfortable in the presence of. . .
Whenever he saw his superior coming he. . .

The assumption underlying sentence completion tasks is that the way an individual completes the sentences is a reflection of his attitude. In the two examples above, subjects favorable toward America and/or Russia would be more likely to complete the sentences in favorable ways than subjects who have unfavorable attitudes. Likewise the completion of the statements concerning superiors and subordinates would be completed in ways which are consistent with the individual's attitude. In both studies cited above the results supported this assumption.

Still another indirect approach is to present individuals with pictures of other people and ask them to respond to what is presumably happening in the picture. For example, in a study of attitudes toward physicians one might present a series of pictures portraying physicians engaged in a variety of activities. Subjects might be asked to describe the setting, the activities, and a probable outcome, or they might be asked to provide dialogue such as the probable response of a patient to a physician who is saying, "I can't help you if you don't follow my orders." Again, it is assumed that the response of the subject to the task reflects the subject's attitudes.

As with direct approaches and the approaches discussed below, there are both advantages and disadvantages to the use of indirect ways of measuring attitudes. The advantages claimed for

indirect approaches are as follows: (1) they encourage in respondents a state of freedom and spontaneity of expression; (2) they can tap a person's attitudes on issues that they cannot easily evaluate or describe their motivations or feelings; (3) they are particularly useful when they are employed on topics on which respondents may hesitate to express their opinions directly for fear of disapproval by the investigator (a major problem with direct approaches); (4) they may be the only means available, e.g., when respondents are likely to consider direct questions as unwarranted invasion of privacy or to find them threatening for some other reason.

While many of the indirect measures are highly ingenious, an investigator must consider their disadvantages before deciding to use one of them. The main disadvantages are: (1) they usually involve at least some degree of deception and occasionally some invasion of property, since individuals are induced to respond under some pretext other than the investigator's true interest and since they are encouraged to reveal matters that they might perhaps wish to conceal; and (2) very few, if any, of these measures have been subjected to any extensive evaluation of either their reliability or validity. That is, investigators employing the same indirect measure often get conflicting results, and indirect measures do not correlate very highly, if at all, with other types of measures designed to tap the same attitude. Perhaps because of reliability and validity problems, indirect approaches to studying attitudes are not used very frequently in social psychology.[1]

Physiological measures of attitudes. At the opposite end of the continuum from measures relying on an individual's self reports are those measures relying on physiological responses not subject to conscious control. While the study of such measures depends, of course, on the subject's willingness to cooperate, the results are usually independent of either self knowledge or willingness to report. The usual procedure is as follows: individuals are exposed to the presence of a member of an object group or to pictorial representations in situations involving members of the object group, and involuntary physiological reactions are recorded simultaneously. These measures often involve the galvanic skin response, blood pressure, heart rate, and dilation or constriction of the pupil of the eye. These measures are based on the fact that physiological changes accompany the experience of emotion, and the underlying assumption is that the physiological measures of these changes are indicative of attitudes.

[1] It should be pointed out that these measures are popular in the field of clinical psychology. In fact, many of the measures that we have discussed have been adapted from tests designed for clinical populations. However, even in clinical psychology the evidence for either their reliability or validity is in question.

As an illustration, Rankin and Campbell (1959) employed two experimenters, one white and one black to attach and adjust the electrodes necessary for measurement of the galvanic skin response. Results indicated significantly larger galvanic skin responses when the black experimenter adjusted the electrodes than when the white experimenter did. Similarly, Cooper and his associates (Cooper & Siegal, 1956) found greater galvanic skin responses to the names of negatively valued groups than to those of neutrally valued groups. In addition, they found that galvanic skin responses increased to both complimentary statements about disliked groups and derogatory statements of valued groups. In each case, the underlying assumption was that the changes in physiological arousal was a result of the individual's attitudes.

More recently, there has been mounted an impressive series of studies which indicate that the dilation and constriction of the pupil of the eye is related to an individual's attitudes. Specifically, Hess's evidence indicates that an individual's pupils dilate in response to pleasurable stimuli and constrict in response to unpleasant stimuli. These promising findings, along with the great potential that social scientists often see in physical measures, made this technique quite interesting and even exciting. However, recent systematic research by Woodmansee (1970) has not only failed to replicate Hess's results but has further shown that the pupil of an individual's eye not only dilates to pleasant stimuli but to extremely unpleasant stimuli, e.g., a picture of a filthy toilet in a broken-down bathroom or a picture from a gruesome murder case involving a local coed. Thus, at present the dilation of the pupil of the eye may not be as promising a technique as was originally thought, although it may very well at least index interest and attention.

While physiological measures have the advantage over direct measures that it is more difficult for the subject to take or give false answers and the apparent advantage over indirect measures of being more precise and objective, the disadvantages are also very apparent. First, the obtaining of attitudinal measures is usually restricted to a defined physical setting where the available resources permit proper recording. Second, increases and decreases in physiological arousal cannot be interpreted without knowing what the environmental stimuli are to which the subjects are responding. Third, at least with the physiological measure of dilation and constriction of the pupil of the eye, there is serious concern with respect to interpretability. Fourth, studies employing more than one physiological measure to tap the same attitude often result in one measure indicating a finding that the others do not; when this occurs it raises

questions of exactly what the various physiological indices are measuring.

Notwithstanding these criticisms, physiological measures of attitudes may very well prove to be more reliable and valid in the future. Work by Cook (1968) indicates that subjects who were conditioned to respond favorably to statements concerning the attitudinal object responded favorably in terms of physiological measures to other positive statements, and subjects who were conditioned to respond negatively to the attitudinal object responded negatively to other negative statements. Results of Cook's work are promising, but this technique is not far enough along to warrant any
50 conclusions about its usefulness as an attitude measure.

Observational methods. Another approach to measuring attitudes is to observe an individual interacting with some social object. For example, Mehrabian (1969) has mounted a program of research which indicates that nonverbal behavior is clearly related to attitudes toward another person. In particular, Mehrabian finds that positive attutides are related to assuming closer interpersonal distances, more eye contact, more direct shoulder orientation, and more forward-lean than are negative attitudes. In other words, our nonverbal behaviors are more intimate with those whom we like than with those we do not like.

Another area of research that is making use of observational methods is the field of program evaluation, concerned with measuring the effectiveness of social programs. Public institutions concerned with such topics as health, crime, and education are increasingly being called upon to demonstrate the effectiveness of programs which taxpayers are supporting. For example, Bickman (in press) in evaluating the effectiveness of a mass media campaign designed to encourage the reporting of shoplifters, found that the campaign was effective in communicating and altering an individual's intentions but not in increasing the number of cases that were reported. In other studies appraising the effectiveness of the mass media it has been found that there was little effect on such behaviors as aggression (Feshback & Singer, 1970) or automobile seatbelt use (Robinson, et al., 1973).

Individuals who are concerned with social action research often have employed observational methods. A good example is provided by Saltman (1975). Concern over the implementation of anti-discrimination housing laws led Saltman to audit a number of real estate companies in the Akron area. Saltman sent black and white volunteers to each real estate company. The volunteers kept written accounts of their observations which were then coded to indicate possible forms of discrimination. The results indicated that twelve out of thirteen companies practiced some form of discrimination.

As with the other ways of measuring attitudes, observational methods have their advantages and disadvantages. The advantages of observational methods are: (1) they can tell us a great deal about behavior patterns; (2) they can aid us in the selection of problems and hypotheses; (3) observation may be the only feasible method by which to gather data, e.g., research with children or schizophrenic persons or research concerning how people react to natural disasters; (4) they allow an investigator to record an individual's ongoing behavior as it occurs; thus scientists concerned with how people interact under certain circumstances can observe their behavior under those circumstances.

The disadvantages of observational methods are numerous. First, ethical problems (invasion of privacy) do arise, particularly when individuals are unaware that they are being observed. Second, when people know that their behavior is being observed, the investigator frequently encounters the same problem as with self reports, e.g., subjects alter their behaviors to make themselves look good. Third, it is not always clear whether the observed behavior reflects an underlying disposition (attitude) or whether the behavior is appearing for some other reason, e.g., a behavior may be nearly independent of external patterns of stimulation. Fourth, without the manipulation of variables it is difficult to clearly establish cause and effect relationships.

Despite these advantages observational methods have become increasingly popular over the past few years. As social psychologists have become increasingly interested in ecological psychology, environmental psychology, social action, and program evaluation, observational methods have acquired more respectability then thay had in the past.

Attitudes and the predictions of behavior.

Recall that the underlying rationale for studying attitudes is that what people say is a good predictor of what they will do. Below are a series of summary statements made by authorities who have analyzed and evaluated the numerous studies on the relationship between attitudes and overt behavior.

> Studies on the relations of attitudes and behavior have almost consistently resulted in the conclusion that attitudes are a poor predictor of behavior (Ehrlich, 1969).
>
> Attitude research has long indicated that the person's verbal report of his attitude has a rather low correlation with his actual behavior toward the object of the attitude (McGuire, 1969).

Most researchers have had little success in predicting behavior from attitudes toward ethnic groups (Brigham, 1971).

There is a growing awareness among investigators that attitudes tend to be unrelated to overt behaviors (Fishbein & Ajzen, 1972).

The best known example of the discrepancy between attitudes and behavior came as early as 1934. A social psychologist, LaPiere traveled from coast to coast with a young foreign Chinese couple, stopping at over 250 hotels, autocamps, cafes, and restaurants and receiving normal service in all but one. Six months after the trip, LaPiere mailed to each of these establishments a simple questionnaire which included the question "Will you accept members of the Chinese race in your establishment?" The answers he received were 92% "No," despite the fact that all of these places had, in fact, served his Chinese friends not long before. In other words, the verbal responses were just the exact opposite of the behavioral responses. This state of affairs not only defies intuition and common sense, but it has frustrated and annoyed social psychologists for years.

In attempting to account for the failure of attitudes to predict behavior, social psychologists have identified three factors in addition to an individual not knowing what his attitude is or to lieing. These factors are measurement problems, conflict among attitudes, and situations. Let us briefly discuss each in turn.

Measurement. The typical procedure has been to determine a person's feelings toward a general class of objects (members of the Chinese race) and use this information to predict that person's behavior toward a particular member of the class (a Chinese couple). The more the particular member of the class deviates or differs from the general class the more difficult it becomes to make accurate predictions. In LaPiere's study the Chinese couples that were admitted to the various establishments may have possessed very few, if any, of the characteristics or stereotypes held by the subjects. In fact, by being well-dressed and in the company of an occidental professor, they were almost certainly not much like the image of "a Chinaman" that proprietors intended not to serve.

Conflict among attitudes. People often have more than one attitude toward any object, and the discrepancy between attitudes and behavior often occur because other more dominant attitudes are operating in a particular situation. For example, a physician who is a strong proponent of HMOs may not be willing to speak publicly in favor of them because of an even stronger feeling that physicians should not actively lobby for their own medical interests. The intensity and extremity of attitudes both probably vary somewhat from time to time as a result of recent experience, and an attitude may be strong enough to be dominant at one time but perhaps not on all occasions.

Situations. Perhaps the most important factor accounting for the discrepancy between attitudes and behaviors is the constraints or behavior that exist in any situation. Situational factors are very powerful determinants of behavior. We are not really "free" to behave in any way we might like in just any circumstances. Some of the constraints represent incapabilities of responding in certain ways in certain situations; others represent constraints derived from social expectations and rules. As an example of the first kind of constraint, it has been noted that policemen do not seem to change their behavior very much, even when they know they are being observed, and they often engage in rather undesirable or unprofessional behavior with observers present. One possible explanation that has been posed for such behavior is that the behavioral repertoire of many policemen is quite limited, and they literally cannot behave differently than they do in some situations. Another constraint by inability to respond would be failure to donate to a highly favored charity because of lack of money at the time of solicitation. The kinds of constraints stemming from social conventions are illustrated by the substantial uniformity of behavior in church, the fact that military enlisted men will usually say "Sir" even to officers for whom they have no respect, etc. The difficulties that have been met in identifying consistencies in behavior, accompanied by recognition of the very obvious and substantial importance of situational factors, has led more and more social psychologists to ignore differences between persons and concentrate on situational factors in determining behavior. Whereas 30 years ago the social psychologist's bias was toward individual dispositions, today the bias is toward situational factors.

Conclusion

From what has been presented it is easy and perhaps even logical to conclude that the study of attitudes is a waste of time. Many social psychologists have accepted such a conclusion. While such a conclusion can be partially supported by the empirical data, there is in my estimation still room left for the study of differences between persons in attitudes and related behaviors.

Recently two social psychologists, Bem and Allen (1974) have suggested that part of the problem of identifying consistency in behavior has been to identify the set of behaviors across which consistency is to be expected. For example, if a soldier is asked whether he likes vegetables and then it is discovered that he will not eat rutabaga, kale, acorn squash, or okra, it might not be reasonable to conclude that the soldier does not like vegetables afterall. A better procedure might be to ask first what the soldier considers to be edible vege-

tables and then determine whether he likes those on his list. Similarly, if one wishes to predict whether a person will "cheat" based on self-report of disposition to cheat, it would be a good idea to find out from the subject just what he or she considers to be cheating behavior. Moreover, one can find out directly from a person about behavioral consistency. A student might say "I always keep my room neat and tidy, but my car is always a mess." Both those statements might be found to be true, in which case the student could be considered quite consistent in behavior, but not necessarily within fairly arbitrarily defined categories.

If one wanted to follow such an approach in studying the attitudes of the public toward a rescue service, one would want first to find out what services the respondent believed were provided and what the important factors were in the provision of such services. It might then be determined that the person was consistently pleased with response times and with the technical quality of the services but dissatisfied with the demeanor of ambulance attendants while handling lower class and indigent victims.

This approach is promising, but it is too early to make a definitive judgment on its value. However, it should be clear that the assumption of consistency of responses toward social objects has been given up, and social psychologists are now looking at a person's feelings toward a specific object in a specific situation, and then observing for the corresponding behaviors.

In summary, the social psychologist's assumption that attitudes lead to a pattern of consistent responses (particularly consistent overt behaviors) toward a social object cannot be supported by the existing empirical data. Rather, an individual's behavior seems to be affected by conflicting attitudes as well as situational factors. The most promising approach appears to be more specificity in the questions that are asked so as to be able to predict when an individual's dispositions will lead to consistent or inconsistent behaviors.

References

Allport, G.W. Attitudes. In C. Murchison (Ed.), *A handbook of social psychology*. Clark Univ. Press, 1935.

Bem, D.J. & Allen, A. On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, 1974, *81*, 506–520.

Berkowitz, L. *Social psychology*. Glenview, Ill.: Scott, Foresman, 1972.

Bickman, L. Bystander intervention in a crime: The effect of a mass media campaign. *Journal of Applied Social Psychology*, in press.

Brigham, J.C. Ethnic stereotypes. *Psychological Bulletin*, 1971, *76*, 15–38.

Burwen, L.S., Campbell, D.T., & Kidd, J. The use of a sentence completion test in measuring attitudes toward superiors and subordinates. *Journal of Applied Psychology*, 1956, *40*, 248–250.

Cook, S.W. Studies of attitudes and attitude measurement (Mimeograph). AFOSR Technical Report, 1968, Boulder: Institute of Behavioral Science, Univ. of Colorado.

Cooper, J.B. & Siegel, H.E. The galvanic skin response as a measure of emotion in prejudice. *Journal of Psychology*, 1956, *42*, 149–155.

Ehrlich, H.J. Attitudes, behavior, and the intervening variables. *American Sociologist*, 1969, *4*, 29–34.

Feshback, S. & Singer, R.D. *Television and aggression*. San Francisco: Jossey-Bass, 1970.

Fishbein, M. & Ajzen, I. Attitudes and opinions. *Annual Review of Psychology*, 1972, *23*, 487–544.

Haire, M. Projective techniques in marketing research. *Journal of Marketing*, 1950, *14*, 649–656.

Kerr, M. An experimental investigation of national stereotypes. *Sociological Review*, 1943, *35*, 37–45.

LaPiere, R.T. Attitudes vs. actions. *Social Forces*, 1934, *14*, 230–237.

McGuire, W.J. The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Vol. 3, 2nd Ed.). Reading, Mass.: Addison-Wesley, 1969.

Mehrabian, A. Some referents and measures of nonverbal behavior. *Behavior Research Methods and Instrumentation*, 1969, *1*, 203–207.

Rankin, R.E. & Campbell, D.T. Galvanic skin response to Negro and White experimenters. *Journal of Abnormal and Social Psychology*, 1959, *51*, 30–33.

Robinson, J.P. & Shaver, P.R. *Measures of social psychological attitudes*. Ann Arbor, Mich.: Institute for Social Research, 1969.

Saltman, J. Implementing open housing laws through social action. *Journal of Applied Behavioral Research*, 1975, *11*, 39–61.

Woodmansee, J.J. The pupil response as a measure of social attitudes. In G. Summers (Ed.), *Attitudes measurement*. Chicago: Rand McNally, 1970.

Recruitment, Selection, Training and Supervision of Civilian Observers to Work in Police Patrol Operations Research

William Bieck
Kansas City Missouri Police Department
Operations Resources Unit
Kansas City, Missouri

It seems inevitable that if the quality of performance of emergency medical personnel is to be evaluated in an adequate way, observers are going to have to be deployed at performance sites, whether in vehicles or in ERs. The development and monitoring of a good observer team is no small feat. This paper details the procedures followed by William Bieck, who has had unusual success with an observer study in the police field. His paper also conveys a good bit about the procedures which are necessary in order to achieve a high level of data quality control.

Before proceeding with the topic to be discussed regarding the recruitment, selection, training, and supervision of civilian observers who worked on the Response Time Analysis Study, mention should be made of the study itself in order to provide the listener with sufficient background information to assess the context in which the observers functioned.

The Response Time Analysis Study, a five-year project funded through the National Institute of Law Enforcement and Criminal Justice, the research arm of the Law Enforcement Assistance Administration, is currently being conducted by the Kansas City, Missouri, Police Department, the agency which was the recipient of the grant. The major objective of the study was to analyze relationships between time taken to report crime or request police service, process and dispatch citizen requests, respond to locations from which assistance has been required, and measure probabilities associated with on-scene criminal apprehension, witness availability, victim injury, and citizen satisfaction with police response time. The second objective sought to analyze problems and patterns in crime reporting or requests by citizens for police assistance.

A total of six data collection components were established in order to obtain information necessary to address questions generated by these objectives:

1) *Observer Component* The Observer Component, the focus of this presentation, consisted of nine civilian observers, two females and seven males, who accompanied police officers, involuntarily, for a period of ten months. The observers rode four eight-hour tours per week with police officers assigned to police the city's most active robbery and assault beat-watches. The primary responsibility of each observer was to record times documenting officer dispatch, response and arrival to citizen contact and the location to which the officer had been sent. Additional information concerning locations from which and to which officers had been dispatched and a description of on-scene activities, e.g., completion of an offense report, criminal apprehension, administration of first aid or request for an ambulance or other police specialists, etc., was also obtained.

2) *Tape Content Analysis Component* All calls coming into the Kansas City, Missouri, Police Department that are processed through the communications-dispatch center are recorded on tape. The main purpose of this segment of the study was to record times pertaining to the initial connection between citizens and police dispatchers, crime reporting or service requests by citizens and broadcast and dispatch messages to field officers. Additional information also collected included an analyses of the taped conversations between citizens and dispatchers to identify problems in citizen interactions with dispatchers and dispatcher communications in transmitting assignments to field officers.

3) *Citizen Follow-up Interview Component* Individuals who reported crimes, requested police assistance or were victims of criminal offenses were identified and interviewed in order to obtain information regarding the time at which the crime occurred or was discovered, the length of criminal visibility if a suspect was seen, the location at which the crime occurred, the citizen's activities before the commission of the incident took place, the time taken and problems encountered in reporting the incident to the police and the citizen's satisfaction with police response time and the officer's on-scene activities. Additional data collected included the victim's knowl-

edge, if any, of the suspect involved in the incident together with demographic characteristics of victims and witnesses.

As can be seen from a review of these collection components, information is available to construct a time continuum consisting of intervals which, for example, account for the time taken for a criminal offense to occur, the time taken in reporting the incident to the police, the time taken to process the call through the communications-dispatch center, and the time taken by an officer to respond to and contact the citizen who initiated the mobilization.

The three remaining collection components consisted of a "Test Call" experiment to measure the amount of time required to reach a police dispatcher through the police department's "Crime Alert" telephone number (emergency or police assistance), the department's administrative telephone number, and the Southwestern Bell telephone operator. This information, which was collected between the times of seven and one a.m. seven days a week, was necessary to evaluate the subjective responses given by citizens in reporting crimes or requesting police service.

A "Victim Injury Follow-Up" survey was conducted to determine the degree or extent of seriousness associated with victim injury resulting from crime or other emergency medical incidents.

Finally, an "On-Scene Arrest and Conviction Follow-Up" component was initiated to assess probabilities associated with criminal justice dispositions. Tracking Part I felons through the criminal justice system was considered necessary in order to evaluate the ultimate importance of on-scene arrests as a product of rapid police response given the suspicion that convictions for the same grounds as arrest would be few. Reasons for judicial fallouts are also being obtained.

Having considered the methodological framework in which data were collected, specific attention will now be focused upon the Observer component. The decision to utilize civilian observers on the Response Time Analysis Study was made with disciplined reluctance. Although necessitated by the need to obtain information unavailable through more conventional means, the employment of civilians to accompany police officers during routine tours of patrol presents a multiplicity of challenges even for the most astute administrator with a flair toward research. Problems encountered given the decision to employ civilian observers can be couched under three headings:

1) *Administrative* Once statistical calculations had been computed to determine the number of incidents needed for adequate and representative analysis, an exercise that also predicted the number of observers to be hired, the most salient and immediate concern posed by the establishment of an observer program was cost. Suffice it to say that from the project's inception no provision was provided in the original proposal for an observer component.

Unanticipated cost-of-living salary increases which were triggered by unprecedented inflationary rates served to compound concern for budgetary strain during the fledgling stages of the study. Of tantamount importance, salary increases also escalated fringe benefit payments which are computed at fifteen percent of gross earnings.

As a result of the observer program, additional supervisory, liaison, quality control and clerical staff were also needed to coordinate and disseminate information, maintain service records, follow through on chest x-rays and flu innoculations, issue and complete travel vouchers, insurance forms and time records, secure office space, prepare supply and equipment requisitions, and manage a part-time, non-profit placement service for those confronted with bleak prospects of future employment opportunities once field data collection had been completed.

2) *Managerial* With nine full time civilian observers, one overall collection supervisor, one observer supervisor, one liaison officer and one quality control clerk, considerable effort was directed toward establishing lines of communication and delineating areas of responsibility.

The observers were given their own field quarters which contained an office for their supervisor, a conference room and a small but functional message center. Although distance *per se* created ripples of alienation among observers toward the administrative and analysis staff, who were located in the central business district adjacent to police headquarters, separate office facilities were more convenient, being strategically located between division stations where field tours commenced, and provided the observer supervisor with sufficient latitude to acquire a working knowledge of each observer's values, expectations, aspirations, and idiosyncrasies. Concessions were made by the observer's supervisor regarding alterations in scheduling so that exceptions could be made to accommodate those wishing to pursue course work at local colleges and universities. Coordination of training sessions, where observers were required to provide assistance in instrumentation construction and modification, deployment scheduling and control of rumor and innuendo, which surfaces as an incessant problem whenever civilians and sworn personnel are forced to work together,

consumed major attention in addressing managerial issues.

3) *Methodological* In general utilization of trained observers is indicative of the state of the art in which research is being conducted. Observer components exemplify admission that little is known about the subject to be researched. It also suggests that the nature of the investigation is exploratory and descriptive rather than experimental; the latter of which can usually anticipate extraneous variation and hence control, a concept central to scientific inquiry, hold constant or account for influences which might affect the relationships being tested.

Although the utilization of trained observers to collect data on research projects is elementary given its methodological niche vis-a-vis more sophisticated techniques used in elaborate research designs, problems associated with the administration and management of such endeavors are extremely complex. Without pursuing an epistemological tangent regarding the historicity of science, what science is and is not, suffice it to say that two methodological limitations are inherent in observer data collection procedures: 1) *Control Effect* Control effect refers to the change or influence the observer creates by his own presence in the situation he is studying. In more concrete terms, observers riding with police officers who are aware of the observers responsibility to obtain information pertaining to response times might be inclined to drive faster (or slower) in order to impress a novice civilian. Furthermore, officers might feel compelled, knowing that they are being observed, to be more thorough in conduction of on-scene activities, e.g., report taking, processing evidence, etc.; and 2) *Biased-viewpoint Effect* This concept describes the potential for an observer to become emotionally consumed into the situation under investigation thereby militating against his objectivity. An observer might be positively coopted by a patrolman in terms of fabricating data that would place the officer in an unfavorable light or become negative toward policemen and the manner in which calls are handled.

As can be adduced from this discussion, either limitation, unless checked, will lead to serious distortion in data collection and analysis.

Having reviewed the setting in which the observer component functioned and problematic considerations generated by the decision to establish an observer program, it is time to proceed to the business of recruitment, selection, training and supervision of observers.

The qualities necessary for a good observer were not easily defined. The role demanded a person with a complex and sometimes inconsistent set of attributes. A good observer would have to face and handle many ambiguities inherent in police-citizen encounters, requiring him to have considerable adaptability to a broad range of situations. Those situations would vacillate between extreme boredom and intense stress. In addition the role would require an unobtrusive individual who could passively blend into any setting, yet actively collect pertinent and accurate project data. Other characteristics such as good judgment, dependability and honesty would also be necessary to insure systematic observations and qualitative data. Since all observers would be contract employees of the Kansas City, Missouri, Police Department, they would have to pass a thorough background investigation.

Initially, it was decided that only male candidates would be recruited as observers, the rationale being that a female observer functioning in a predominately male line of work would introduce an element of bias to both police officers and citizens by producing expectations for which it would be difficult to control. The easiest role for a civilian observer to acclimate in the police-citizen milieu was either that of a plain-clothes detective or a police recruit; it was considered problematic to cast a female in either role. Since no empirically tested data were available to support such a position, the legal obligation of the police department to be non-discriminatory in its hiring practices (the study was also federally funded) resulted in the position being opened to both sexes.

Initial concern about acceptance of female observers was borne out somewhat during the first weeks of field observation. One incident involved a woman who had called the police regarding a destruction of property complaint. When contacted later by a telephone interviewer she said the officer had arrived late on the scene (he was accompanied by a female observer) and she had assumed he had picked up his girlfriend prior to responding to the call. In another police oriented study involving observers in Rochester, New York, citizens' complaints were so frequent that specially designed blazers had to be worn by all females while conducting their field work.

To mitigate against role conflict between officers and civilians on the RTAS, all observers were required to display department identification which consists of a personal photograph captioned "POLICE—CIVILIAN EMPLOYEE." After instituting that procedure citizen complaints abated.

The only specific criteria first required for application for an observer position was a minimum age limit of 21 and completion of high school degree requirements. It was later learned that the minimum age stipulation was not a department requirement; the lower age limit being

17 with parental concent to work. Minimal application criteria were established because of the lack of evidence regarding the most suitable background for observer candidates. In fact, employment criteria were so general that almost anyone might qualify for the position.

The most immediate market for qualified candidates at first appeared to be local Associate and Baccalaureate Degree programs. As a result all colleges and universities within a sixty mile radius of Kansas City having a liberal arts or criminal justice major were contacted. If the institution had a placement service, it too was contacted. Individuals involved in hiring civilians for the police department were also advised of the observer openings. The initial requests for applicants resulted in only fifteen persons applying for the nine position openings.

After initially receiving a poor response, recruitment efforts were accelerated and expanded to include out-of-state institutions. The Job Information Center at Sam Houston State University in Huntsville, Texas, was contacted. This school maintained several hundred resumes on eligible candidates in the criminal justice field. Northeastern University's job placement advisors for the College of Criminal Justice in Massachusetts were also notified of the openings. Finally the positions were advertised for two consecutive Sundays in The Kansas City Star, the metropolitan area's major newspaper.

The second round of inquiries, including the newspaper advertisements, brought an improved response. Over 200 inquiries were received, and of those, 176 agreed to submit resumes. A total of 104 resumes were finally received; sixty-nine percent from males and thirty-one percent from females.

A revision of the project timetable to resolve research design issues resulted in a two-month delay of interviews for the observers. During that period, sixteen applicants found other jobs, four moved away, four changed their minds, three withdrew citing "bad hours" as the cause (normally observer shifts ran from 4:00 p.m. to midnight), and three others did not attend their scheduled interviews. The remaining dropouts were those recruited from Sam Houston and Northeastern Universities.

Originally, two members of the project staff had planned to travel to Texas and Massachusetts to screen prospective candidates. However, the two-month delay resulted in a diminished list of out-of-state applicants. Travel costs could no longer be justified given the number of candidates remaining, and after being advised that they would have to travel to Kansas City on their own expense (federally funded grants prohibit payment for relocation to new jobs), they declined further consideration.

A total of fifty interviews were finally held with thirty-eight male and twelve female candidates. The selection process involved three basic phases: 1) Personal interviews; 2) Field evaluation; and 3) A battery of short tests combined with a brief open-ended interview. Each applicant had to successfully complete all three stages in order to be eligible for final selection. Each phase was designed to examine particular attributes needed in a "good" observer. Characteristics deemed desirable for competent observation were evaluated in at least one of the three phases, and most were evaluated in a second or third phase supplying a cross-reference indication of ability.

The initial phase involved one police officer and one civilian interviewer questioning each candidate for approximately an hour. Prior to the interview the applicant was asked to print his name, age, height, weight, and telephone number on the cover sheet of an interview form. This provided the interviewers with an indication of the applicant's ability to print letters and numbers legibly, an important factor in the coding of survey data forms, especially in anticipating that raw data would be obtained in moving police vehicles.

The interview began with a general explanation of the study and a job description. The candidate was then asked a series of questions regarding his career objectives, work experiences, educational background, and general interest and aptitude for the observer position. An ambiguous problem situation was described by the interviewers, and the applicant was asked to discuss it. Responses indicating rigid or extreme value orientations on behalf of a candidate were considered undesirable and potentially problematic for the observer role.

At the conclusion of the interview, each interviewer completed a rating form ranking the candidate's listening and communication skills, work experiences and general appearance as it applied to the role of an observer. Preference was given to applicants with a college degree, experience in applied research and a demonstrated interest in police operations. Such qualifications were expected to facilitate individual training for the observer role and improve the likelihood of qualitative data being collected. A summary rating on a scale from one to five was computed for each candidate. After all interviews were completed, the interviewers received the rating instruments to reevaluate any prior rejections and make final first-phase decisions. A unanimous "no vote" was required for an applicant to be rejected. Having completed screening during the first phase, twenty-five applicants remained eligible for the second phase (although twenty-five applicants were designated for the second phase, three withdrew, having accepted other positions, and one moved away).

The interview teams involved in the selection process included two civilian employees of the department who had considerable experience as observers in police patrol operations and two police sergeants who had an extensive knowledge of field operations. One of the sergeants has been previously selected to supervise the field observers during training, pretesting of field instrumentation, and data collection. Several police officers and a department operations analyst conducted interviews when the regular interviewers were not available.

It was intended that at least one civilian and one uniformed officer would participate in the interview sessions, however, scheduling conflicts resulted in thirty percent of the interviews being conducted solely by sworn or civilian personnel. Of the fifty interviews completed, thirty-five involved both civilian and sworn interviewers, twelve involved strictly sworn interviewers and three included only civilian interviewers.

The second stage of the selection process required that observer applicants accompany police officers during routine patrol tours for a minimum of sixteen hours (normally twenty-four), after which evaluations were made by a preselected group of police officers. Candidates were given minimal instructions on how to behave and were expected to improvise in some situations. The evaluating officers were chosen by the sworn members of the interview teams to represent a variety of personalities and methods of employing police procedures in an attempt to expose each candidate to a variety of policy styles, which were anticipated to be encountered during the fifteen months of field observations. Officers rated the prospective observers on the basis of compatibility (the major consideration), job interest, supervisability, courage, and inconspicuousness on calls. At the conclusion of each tour, a police sergeant conferred with the evaluating officer and compiled a ranking of those candidates evaluated by that officer. This process allowed an officer to reassess his earlier ratings given the broader field of reference he had developed. Only those candidates who received acceptable ratings from all officers with whom they rode were selected for the final phase. Of the twenty-one applicants that took part in the second phase, only twelve qualified for the final phase.

The final phase of the selection process included a battery of paper and pencil tests and an open-ended interview with the interview team. The first exercise was a picture-recall test which is used by the Regional Center for Criminal Justice to determine police officers' ability to observe details at a crime scene. The second exercise was a digit-symbol drill testing the candidates' dexterity and ability to print legibly. The final test, one developed by the Shipley Institute, provided an indicant of the applicants' abstract reasoning ability and I.Q. level.

Once the tests had been completed, the observer candidates were again screened by interviewers in an open-ended interview. This provided interviewers, who had not previously seen some of the applicants, with a complete review of the final candidate field. The applicants were then ranked according to their scores in the second and third phases. These scores were considered with the personal evaluations of the interviewers in the final phase, and nine candidates were selected.

The nine individuals chosen constituted a diverse group. The oldest member was a thirty-six year-old female with a Master's Degree in Public Administration, who, incidently, resigned after the first week of training to accept a position as director of a youth service agency. Her replacement was a twenty-eight year-old male who had been designated as an alternate from the final field of twelve. The youngest observer was a twenty year-old male with a high school diploma.

Of the nine observers selected, seven had Baccalaureate Degrees of which three had also completed Masters Degrees. The average age of the observers was twenty-seven years. Final selection revealed that one-third of those initially selected was female. The group represented a variety of work experiences which included a correctional officer, a weather observer, a personnel technician, a psychiatric aide, a clerk typist and a research assistant.

Given the fact that those qualities which describe a "good" observer could not be defined at the outset, a meticulous selection process does not guarantee a successful observer program. Once selected observers must be trained and then supervised throughout the entirety of data collection.

Observer training on the Response Time Analysis Study sought to achieve two objectives. First, it was expected to provide observers with a thorough understanding of police operations. This was considered necessary given the length of time data was projected to be collected and the realization that civilian observers would be riding with police officers in the highest crime areas of the city. Secondly, training was designed to insure that observers received a complete orientation regarding research methodology utilized on the study together with instruction concerning all collection and quality control components. Through a combined review of the occupation to be researched and the nature of the research to be undertaken, the observer gained a more complete understanding of his work and the responsibilities of sworn officers.

Before addressing specific aspects of the training format, a brief but important sidetrack is warranted. Having made decisions to establish an ob-

server program, the number of observers to be employed and the methods by which candidates were identified, recruited and selected, the subject of the kind of observational technique needed on the Response Time Analysis Study was discussed. Hopefully, the following labels are self explanatory, but there are at least four types of observer alternatives: 1) Participant Observer; 2) Observer Participant; 3) Complete Observer; and 4) Complete Participant. Although differences among these methods vary in degree, they also vary in kind. The distinction between a Complete Participant and a Complete Observer is absolute. Perhaps of interest to the layman is the fact that these methods are also utilized by individuals outside the research community. For example, an undercover narcotics agent might wish to infiltrate a drug traffic operation in order to secure evidence. His "cover" or "front" must appear legitimate to his adversaries before admission and then participation in the group is permitted.

In short, unequivocal guidelines were established at the outset of training to define the observers' role as "complete observers." Their mission was first and foremost to *collect data!*

Actual training involved a collaborative effort among policemen, civilian researchers and project consultants. Training units on patrol operations, street and field procedures, first aid self-defense and other aspects of police work were provided by the Field Operations Supervisor, who was in charge of the observing, with assistance from a retired police sergeant, who was the project's field liaison officer. Observers were given a tour of specialized units within the department, e.g., K-9, helicopters, traffic, etc., and received instruction on the operations and objectives of those units from member representatives. A seminar on epistemology, science and research methodology was conducted by the Principal Analyst, a former Assistant Professor of Sociology. Sessions on field data collection techniques and instrumentation development were delivered by an Operations Analyst, who had conducted field observations for over a year while employed on the Preventative Patrol Experiment, a study conducted in Kansas City which was funded through the Police Foundation. A special session dealing with the potential of observer co-optation and the concept of "going native" was presented by Dr. Albert Reiss, a Professor of Sociology from Yale University. Dr. Reiss had had considerable experience in directing observer programs in other police departments. Finally, an orientation to the department's overall research and programmatic activities was provided by the unit commander of the Operations Resource Unit, an operational planning agency responsible for organizational development and applied research efforts within the department.

Training topics included a project orientation (16 hours), rules and regulations (3 hours), a department orientation (18 hours), police work (42 hours), research methodology (16 hours), instrumentation development (76 hours), and field work (72 hours). Over sixty percent of the program was focused on instrumentation development and field work. In sum the observer training program consisted of 243 hours of instruction, field tours, seminars, and discussions.

In the initial training session a complete review of the Response Time Analysis Study was presented. Its origin, objectives, methodology and potential implications for the Kansas City, Missouri, Police Department and the law enforcement community were discussed. Emphasis was given to the necessity of systematic and honest collection and recording of observations.

A discussion was held on the rules and regulations of the department as they applied to civilian employees. This included a review of the legal rights and obligations of department members. Specific emphasis was given to the following administrative guidelines regarding study personnel which was formulated by Response Time Analysis Study staff and then approved by the Commander of the Operations Bureau:

1. Project staff shall treat survey data, incidental observations, and official departmental business as confidential unless release is authorized by the Project Director.
2. Survey data and other information incidental to project objectives will be provided to the department for matters involving criminal investigations.
3. Departmental personnel involved in processing and having access to project data shall refrain from discussion of such information, regardless of how incidental, unless authorized to do so by the Project Director.
4. Sworn personnel accompanied by project staff will remain anonymous to project reports. Information obtained from communications and field operations will be statistically tabulated in aggregate form for analytical purposes only.
5. Civilian study personnel are not permitted to assist sworn officers unless dire necessity indicates such behavior is appropriate. However, study personnel are required to provide assistance, i.e., physical or other reasonable actions, to sworn personnel upon command, or when it is obvious and apparent that specific situations dictate such actions.
6. Survey data and other extraneous information obtained by project staff, i.e., incidental observations, etc., will be exempt from departmental use for disciplinary purposes against sworn personnel, except for those incidents involving criminal conduct. Proj-

ect employees are required to report both illegal actions and incidents of questionable legality to the Field Operations Supervisor.

These guidelines were distributed in an Operations Bureau Memorandum to all members of the Kansas City, Missouri, Police Department. It specified a code of conduct distinguishing the project staff from other department members and insured pledges of confidentiality would be honored.

Observer orientation also included several hours of instruction regarding the operations and organization of the police department which included a tour of police headquarters, various specialized units, the county jail, and the municiple and criminal courts. Presentations were made on the organizational structure of the department, allocation of resources, operations of division stations, jurisdictional areas delineating police responsibilities, and the criminal justice system. This orientation provided observers with a basic understanding of the organization being researched and its relationship to other judicial systems.

One of the major training components, which required over forty hours of instruction, was police work itself. This segment focused on police training and field procedures applicable to police patrol. An introduction of police work was presented in a training film entitled "Law and Order" which depicted different aspects of police work in Kansas City, Missouri. Instruction was given in self-defense, first aid, equipment usage, department procedures for handling specific incidents and on-scene criminal investigations. Observers were also familiarized with the uniform crime reporting policy, department reporting forms, report writing procedures and beats targeted for field observations.

In the methods section the observers received an introduction to research methodology and field data collection techniques. Observers received instruction in role playing and observational field procedures, which could be utilized in reducing observer bias and optimizing data collection workloads. Additionally, discussions were held on appropriate attire and acceptable equipment which would offer the most unobtrusive appearance for observers in police-citizen encounters.

Approximately thirty percent of the training program focused on instrumentation development. Initially, a review of the observational instrument was presented in the context of project objectives. Subsequent meetings examined instrument items, operationalization of terms, refining skip patterns and simulating encounters to be coded. Extensive sessions were conducted throughout the training and pretest periods in order to review and revise the field instruments and problems identified in collection of data. To assist in clarifying some of the more complex terms and instrument items, observers were divided into groups of three to research and recommend concept definitions and syntax of the items for the observer survey form.

Field work was conducted throughout the training and pretest periods. Observers initially rode in police cruisers in different parts of the city for a general orientation of patrol and to become familiar with dispatch communications procedures, policies, and communication jargon. They were instructed not to take notes, but simply to act as observers of mundane police activities. This allowed them to become familiar with police work and the officers without being burdened by data collection. Once a degree of familiarity and credibility was established, some limited data were collected to orient both the observer and the officer to what would become the observers normal work routine. After instruments were constructed and equipment acquired, each observer was accompanied by the principal analyst and the operations analyst in charge of establishing the observed component and field instruments for a complete tour of duty during which time measurement differences were monitored and field collection techniques discussed.

Throughout the training period a continuing dialogue on the need for qualitative data and the honest reporting of mistakes was encouraged. Meetings with the Kansas City, Missouri, Chief of Police, the Response Time Analysis Study Project Director, several consultants and staff were held to emphasize the need for maintaining a high standard of integrity in conducting field observations. This theme continued to be emphasized throughout the pretest and actual collection phases of the study. In order to document the extent to which observers conformed to project guidelines, however, adequate supervision needed to be provided and quality control checks implemented.

A sector sergeant from the Kansas City, Missouri, Police Department had been selected by the Project Director to supervise the observer component following futile efforts to solicit a person who met the qualifications that had been defined for the position. With nine years of street experinece, the rank of sergeant and thorough familiarity with police operations and department policy, it was reasoned that novice observers would find it extremely difficult to fabricate data pertaining to response times and on-scene police activities.

Training emphasis of the Field Operations Supervisor was placed on research methodology and the study objectives. He was familiarized with the study components, available literature pertaining to previous research on response time and other observer programs. Briefings on supervisory and observer responsibilities, quality control systems and department liaison were conducted with project consultants and study staff. Most of the training, however, resulted from first-hand on-

the-job exposure in working on the observer selection, training, and pretesting phases of the study.

Once study objectives had been articulated and a methodology developed, department cooperation and support had to be secured. This required those individuals in the department most affected (or threatened) by the study to receive a thorough orientation of project plans. Given the hierarchical structure of the police department, all levels of the organization had to be informed. Since the areas targeted for observation included all three divisions, commanders, desk sergeants, sector sergeants and patrol officers from each division were familiarized with the study.

60

There were many problems which could be anticipated in the conduction of this kind of research. For example, the tendency of police officers to be suspicious could result in observers being labeled as spies. In addition there was some danger of information distortion as it filtered through the different organizational levels of the department. Finally, observers once accepted in the field setting might be pressured to take a more active participation in police work.

To minimize these and other concerns a retired Kansas City, Missouri, police sergeant was hired as an assistant Field Operations Supervisor to help maintain sound working relationships between project staff and operational personnel. He was well qualified to act in the liaison capacity having served in the department's operations division for nineteen years. During his tenure on the department he had established a reputation of dependability and personal integrity.

The assistant field supervisor's primary duties included:

1) Meeting with and orienting district officers to the project and discussing with them any problems resulting from the observational program.
2) Familiarizing desk sergeants with the Response Time Analysis Study and observer allocation needs.
3) Informing pertinent command staff of study objectives, project progress and potential implications of research findings.
4) Interviewing field sergeants to formulate observer procedures when riding with officers and to ensure that police personnel were not discriminantly assigned due to observer deployment.
5) Maintaining a general knowledge about the organizational environment and receptivity to various project related procedures.
6) Monitoring personnel changes of district officers assigned to the target areas and familiarizing newly assigned personnel with the study.

The assistant Field Operations Supervisor was also required to submit a quarterly report to the Field Operations Supervisor regarding feedback from police officers indicating any problems encountered as a result of observer data collection procedures or the conduct of the observers themselves.

The following quality control checks were established and monitored during field data collection:

1) All data submitted to the Field Operations Supervisor had to be reviewed beforehand and initialed by each observer to insure its completeness and accuracy.
2) Police officer activity sheets were checked against the observer's log of eligible incidents to insure that data were collected on each call.
3) Wrist watches worn by the observers were synchronized every two weeks with the master recorder located in the communications-dispatch center. Variations of time differences were recorded in order to identify faulty time pieces. In addition periodic battery inspections of watch modules were made to avoid malfunctions.
4) Chronological logs were developed to monitor disciplinary, managerial, administrative, research, and equipment problems. Information was scrutinized to identify if problems clustered in specific areas, were randomly dispersed among observers or were manifest to specific individuals.

Once observer instruments had been checked by the Field Supervisors, all data were forwarded to the Quality Control Clerk who was stationed in the downtown administrative and analysis office. The primary responsibility of this person was to catalogue field forms by precoded number and disseminate them to the appropriate collection component supervisor.

Now that observer data collection has been completed for over eight months, evidence indicates that the observer component experienced minimal problems. Exit interviews of observers before their departure substantiates earlier supervisory and consultant reports regarding the quality of data collected.

The "control effect" discussed earlier appears to have diminished as a major limitation inherent in this observational research given the number of other factors which also influenced the officers' performance while data were being collected. The "biased-viewpoint effect," which signaled the danger of an observer becoming coopted, was checked almost totally from the outset by the observer deployment matrix which required every observer to rotate beat-watches following each week of data collection. Frequent meetings between the project's liaison officer and police officers also helped reduce the chance of this problem surfacing.

Two suggestions for EMS administrators are warranted following experiences obtained from the project just reviewed and consultation with other researchers and administrators. First, there is absolutely no reason to feel apologetic or defensive regarding research possibilities within your own agencies. So little is known about even the most elementary assumptions in urban emergency services that researchers are often themselves embarrassed. If research contracts are negotiated or grants developed, make sure provision is made for a special liaison consultant to evaluate the work being conducted for your own benefit. This person could be recruited locally and would provide valuable insight into interpretation of project findings and assessment of implications. Secondly, in order to respond to officials in other administrative positions and the press, allow sufficient funding to establish an implications committee which would explore consideration for new programs in the event that shallow results were reported. All too often researchers have told public administrators what doesn't work without suggesting constructive alternatives.

Developing Indicators of Program Effectiveness: A Process

George L. Kelling
Police Foundation
Washington, D.C.

62

One of the requirements for evaluating any program is that adequate measures of program effectiveness be devised. Although many programs have explicit outcome measures such as might be the case for a communications system designed to decrease response time, it is often necessary to devise "indicators" of effectiveness which are somewhat indirect or removed by some steps of inference from the effects actually intended. In this paper Kelling describes some of the problems in developing indicators and points to ways of maximizing their validity.

The development of indicators of program effectiveness is tricky and important business.

Perhaps the easiest way I can make this point is to give some examples from policing, the area in which I do my own research and evaluations. I will present, and discuss three examples. I will then close by describing the process which I feel is necessary to develop indicators for evaluations.

One of the problems in policing about which there has been recent concern, has been the problem of police brutality. Many programs have developed to deal with this problem. Solutions include, citizen review boards, peer review panels, training, retraining, enlightened disciplinary procedures, higher education, psychological counselling, etc. An indicator of police brutality is the number of complaints filed against police officers. But, I know of a city where police-citizen complaint centers advertise their location, where citizens are encouraged to complain if they are not satisfied with services, where citizens' complaints are processed rapidly and continuously, and citizens are kept informed of the procedures and actions that the department takes. I know of another city where citizens can't locate where they are to complain, are discouraged from complaining, and are never informed of the outcome of their complaints. The first city has many complaints. The second city has few.

The point in this example is relatively simple. The meaning of indicators is relative to their context. With all deference to Gertrude Stein, "A complaint is not a complaint, is not a complaint, is not a complaint." The same thing could be said of arrests, crime statistics, and a host of other indicators.

In this example it is clear that the activities of one organization have encouraged citizens to complain and made the complaint process so accessible that it is not unlikely that they will accumulate many more complaints than the department which discourages complaints and makes complaint locations inaccessible. The number of complaints then, may not be an indicator of brutality, but rather an indicator of the success of a complaint processing system. It *may* also be an indicator of brutality, but that may be extremely difficult to discern.

Likewise, it would be possible that a police department could, with great fanfare and publicity, embark on a program to reduce complaints through training, recruitment, discipline, etc. That program, attended by publicity, could call attention to police behavior to persons who, in the past, simply gave it no attention ("What the hell, so police do thump once in a while"), thus modifying public expectation of behavior, which in turn would lead to *increases* in complaints. Those increases could occur in spite of the fact that officer behavior improves. It is conceivable then that an increase in complaints could indicate a change in citizen expectations rather than officer behavior.

Let me give yet another example in this area. We know that there is a great gap between *actual* levels of crime and reported crime. How large that gap is, varies from place to place and from crime to crime, but generally it is known that 50% of crime goes unreported.

Let us suppose that a department goes into a vigorous anticrime program which includes crime specific strategies, eliciting more information from citizens, and improving police-citizen relations. Let us further suppose that in the process of conducting this program the police manage significantly to affect the public perception of their effectiveness. It is not unlikely that many citizens who have failed to report crimes because they have felt the police could not or would not do anything about it (remember that 50% of crimes go unreported) would start to *report* crimes which they would not

have in the past. If reported crime is an indicator of effectiveness, reported crime could go up and the program could be viewed as a failure. In fact, the increase in reported crime could mean that the department had been successful in improving public confidence in their performance. (Rape is a good example. Rape is seriously under reported. Rape victims are, more and more, being encouraged to report rapes and, in response to public pressures, police departments are improving the quality of their handling of rape victims. It is conceivable that *reported* rapes will increase but that does not mean that actual rapes have. They may have, may not have, or may have stayed the same. Increase in reported rape statistics can be the result of changes in public mores and improved police procedures.)

One more example. One of my colleagues, Mr. John Murphy of the Police Foundation, has been examining the issue of arrest productivity in police departments. (Arrests have been one of the historical measures of police productivity). As he went from department to department he found tremendous disparity in the numbers of arrests that officers made which seemed to have no relationship to reported crime or victimization levels. That led him to the second question. "What does an arrest mean?" After months of immersing himself in that data, he has identified the myriad of factors that can be, and are, related to arrests. (Organizational factors, police style factors, reward factors, neighborhood factors, actual crime factors, definition of crime factors, court factors, etc., etc., etc.). The point is that the meaning of arrest, as with all indicators, is tied into a variety of contextual issues. To know what arrest, complaint, crime, morale, job satisfaction, etc. indicators mean, each must be seen within a context. If the context is not understood, indicators can be interpreted as meaning one thing when, in fact, they mean something diametrically opposite.

> I know of a proposed evaluation of police services in which two principle indicators of police performance are response time (how long it takes for a police vehicle to respond to a call for service), and police passings (the number of times a police car passes a particular point). The assumptions are that if a police vehicle responds rapidly, criminals will be apprehended or deterred and citizens more satisfied, and that if a police car passes a particular point often, criminals will be deterred and citizens made to feel more safe. It seems logical that both response time and passings are indicators of police performance. Yet while that appears logical, there is no empirical evidence that either fast response time or number of passes accomplishes anything.

The theories have been that rapid response time and passings can lead to crime reduction, apprehension, and citizen safety. But those have remained, at least until very recently, unexamined theories, and unexamined assumptions.

The development of these two "indicators" of patrol effectiveness has been an interesting phenomenon in policing. Measuring patrol effectiveness has been a particularly thorny problem in policing since so much important police activity (public service) has been inappropriately relegated to second level importance and crime related activities (crime related functions account for, at the most, 20% of police time) have assumed exaggerated importance. The combination of the excitement of the criminally related activities and the "Kojak Syndrome" has led both the police and students to the police (read that as researchers and evaluators) to virtually ignore public service functions and indicators in evaluations of the police. Coupled with that functional bias, and the difficulty of measuring effectiveness, response time and passings (technically but expensively measurable) based *only on theory and logic,* have come to be substituted for actual goals. Police *and* evaluators have become willing to assume that if response time is low and passings often, that that, in itself indicates success. In point of fact, it indicates only that response is low and passings often. Means have been substituted for goals.

One has to be careful not to be too harsh about this however. Measuring goal attainment can be extraordinarily difficult. Oftentimes administrations *have* to find *process* (means) indicators to demonstrate their effectiveness since they lack the funds, time and skills necessary for evaluation and, under pressure, they must do as best they can. Likewise it is often the case that as a result of lack of funds or finely developed evaluation methodology, evaluators simply have to settle for process indicators. When that is the case and the theoretical biases and the reliance on means rather than goals are made clear that is acceptable. The mistake occurs when administrators and evaluators come to confuse means and goals. Short response time and many passings, can be achieved, but in achieving those ends, the funds and creative energies are withdrawn from finding techniques which obtain the goals.

> Arrests are oftentimes considered an indication of police performance. The theory is that the more arrests an officer makes, the more crime he is stopping, the more proficient he is as an officer, and the more he is contributing to the solution of a major social problem. Many people agree with that. Labeling theorists argue otherwise. They argue that arrests stigmitize an individual, can create a deviation amplification feedback loop, and make the problem worse for both the individual arrested and society.

With this third example, I am trying to make two points that both evaluators and agency professionals must be extremely clear about. 1) Program

evaulations ought to strive to link theory and practice. 2) There are value ambiguities in many of the goals of social programs.

Regarding the latter point, the long range goal (value) regarding crime in our society seems to be fairly universally agreed upon, that is—to work towards a situation where citizens can live in their homes and in public places with relatively little fear of being victimized. But the interim goals on the way to that broad social goal are not always agreed upon. For some, police are to arrest offenders and present them for rapid processing. For others, the police are to divert offenders, especially young offenders, from the criminal justice system. Cost effectiveness and cost/benefit models tend not to emphasize the function that values have in determining program goals or their measurement. As complicated as the cost/benefit and cost effective equations are, they are only meaningful when placed in the context of values.

Dealing with social problems involves delicate value and norm decisions. No doubt it would be possible to deal with many problems more effectively if we were not restrained by values and standards. Crime is an excellent example. Concern for issues like privacy, due process, and humane handling of individuals restrains organizations as they work towards their goals. The point is that agency personnel have to context their goals within the broad values of society. Goals are always values or contribute to values. I am now asserting this as more than an abstract truism. It is an important fact that politicians seem often to be more aware of than we—as they ignore our cost benefit calculations.

Further, theories play an important function in our work. As evaluators and agency professionals work together to establish goals and indicators of those goals it is important that they understand that all social practices have, or at least ought to have, explicit theoretical bases and that the evaluation of program outcomes should be a test of theory. While some of our evaluation activities are mundane and tedious, others call for us to return rigorously to theory and attempt to understand the relationship of the program evaluated and the theoretical bases of that program (explicit to the agency or not). A program is, or at least ought to be, the operationalization of theory. A critical point in the process of bringing together values, theories and programs is that of establishing explicitly, program goals and indicators. True, this may be a struggle, and true too, it may result in incomplete explanations, but the more evaluators and agency personnel struggle to establish the causal linkages, the more relevant will be their findings. Evaluators, at *best* socialized in theory development, and operating personnel, at *best* socialized in theory application, have rare intellectual opportunities when trying to define "What works?", "How do we know it works?" and finally "Why does it work?".

How then ought we develop indicators? It seems to me the process is at least a three fold one.

In the first place, researchers and evaluators have to develop indicators of program effectiveness through the process of total immersion in agency activities. They cannot sit down in several meetings with agency administrators and expect to know agency or professional goals, skills and practices, and the assumed linkages between them. Agency administrators have *a* point of view but often they are far removed from actual practice. Organization operatives have *a* point of view but that too has its limitations. What the evaluators must do to fully understand practice and goals goes beyond conversation and interviewing. Let me give several examples.

We are now beginning to develop plans to see if it is feasible to do an evaluation of foot patrol in New Jersey (New Jersey provides an interesting site as foot patrol operates in 28 cities and is funded by the state.) In the process of developing the indicators of foot patrol (I must confess that we are also developing hypotheses, working relationships, examining data bases etc., *but* even if we weren't doing the other things we would still have to do the following to develop indicators) We have:

—Met with top officials and administrators
—Met with field commanders
—Met with heads of records units, etc. (to see if data are available, how much it will cost to access, and how much has to be generated over and above that which is available).
—Met with a group of supervisors and administrations to discuss what foot patrol is to accomplish and how we can tell if it is accomplished.
—Met with a group of patrol officers to discuss what foot patrol is to accomplish and how we can tell if it is accomplished.
—Walked foot patrol with patrol officers (so far staff has walked a total of 15 shifts and will probably walk a total of 15 more) in a variety of cities.
—Rode with foot patrol sergeants (so far a total of 5 shifts).
—Formed an advisory group of 2 foot patrol officers and 1 sergeant from each of the 5 departments with which we plan to continue our exploration.
—Asked each of the 5 departments to form a small task force to work with.
—Talked to citizens, including merchants, street people, and local residents about their views about foot patrol.
—Met with state officials in two agencies to discuss with them their perceptions of the goals of the program.

The purpose of all these activities was to educate ourselves to what foot patrol was, what it was to accomplish, what it seems to accomplish, and to hypothesize about the causal linkages between means and goals. (Yes, this *is* terribly time consuming and expensive—I would guess about $40,000 worth of staff time and resources will go into developing an appropriate design and indicators—*not* counting agency staff time—and further it may turn out that after all that time and effort a major evaluation would be so difficult and expensive that only a very modest one would be worth the investment, perhaps one which would cost *less* than the planning itself.) But we believe that only in this immersion can we fully work with agency people to establish a proper design and indicators.

In Kansas City, we worked with a task force of patrol officers and supervisors for a year to develop a design and indicators. That task force also recommended, and the KCPD approved, that two police officers work full time with the evaluators during the entire length of the experiment. (True, the functions of those officers went beyond working with us on indicators and included such things as monitoring the experiment, but throughout the experiment one of their major tasks was to help us understand what data meant. One of them, Charlie Brown, now works full time for the Police Foundation and daily works with non-police researchers and evaluators to help them understand what they are seeing.)

Please understand that I am not saying that all wisdom regarding what data means rests with agency personnel. I very strongly believe that *not* to be the case. They have their own biases, methodologies, and vested interests which keeps them from fully understanding what they see.

Instead, I am suggesting that it is important to develop an interaction between persons deeply involved in research and with those deeply involved in practice. It is out of that interaction that indicators develop. The development of indicators is not a research enterprise alone. It is not a practice enterprise alone. It is a process between carefully trained inquirers and carefully trained practitioners. This process *must* be gone through at some point. If it is not gone through early, it will be struggled through later between antagonists saying "That's not what I do", "That's not what I meant", or, "That's not what it means". If the process is properly gone through, the process results in a contract between evaluators and agency. That contract is called a design, developed by both agency and evaluators.

One last word on this. I am not suggesting this process as *a* way to do it. I am suggesting that it is the *only* way to do it. (Even if the evaluators are doing their first, fifth or twentieth evaluation in a particular agency).

The second aspect of the development of indicators is that the researchers have to return to theoretical and practice literature. Most agency practitioners become fairly removed from the literature of their own field. Most have difficulty keeping up with current research, let alone maintaining their interest in theory development, causal linkages, etc. But that is an important task for researchers and one for which they are extensively trained. The development of indicators is not a mechanical job that can be done independently of the intellectual traditions of a field. As an example, my own feeling is that those who started to use response time and passings as indicators of patrol effectiveness made two mistakes. One was that they confused means with goals. The second was that they simply did not understand the historical traditions of the police. Response time and passings are almost completely related to the crime related functions of the police. (Proponents of these as indicators may argue that response time has broader application but if you read their materials, any other functions of response time are relegated to a distant, distant sound.) The problem is that such an emphasis ignores many of the important historical traditions in policing. This problem of research and evaluations lacking context has been a special problem in policing where few practitioners write, and universities are only starting to begin to do research in policing. (For all practical purposes, *no* research exists on police techniques prior to 1962). Thus researchers carry the responsibility of trying to ground their research (evaluation) in theory. That may be difficult (the Police Foundation accomplishes this partly by having an Evaluation Advisory Group, all of whom are respected academics, whose purpose is to force evaluations to go through the process of trying to tie their work to historical trends and establish the causal and theoretical relationship between findings and practice), and often is exceedingly painful but it is absolutely necessary for a field of practice.

And thirdly, the task of the evaluator as he developed indicators is to help the practitioner context their experiences. In the first point, I emphasized the need for the evaluator to immerse himself in the agency and learn from the agency. Now I am emphasizing the other side of this. It is the obligation of the evaluator to bring to the operating agency the contexts and theoretical traditions discussed above or the evaluator does not just bring the agency technical skills or speak to the agency on its terms, but rather brings a critical capacity both as a result of his/her training and the present state of the literature. He/she conveys to the agency specific research findings and critical analyses of the agencies' program. The evaluator brings these traditions in the form of constant probing and questioning. He/she, by challenging, even irreverently, the present beliefs, can contribute to the learning of the agency. Again, the re-

searcher is an inquirer. The evaluator has to force the practitioner to review his ideas in the context of theory, and history.

Conclusion

I have presented the development of indicators as a process which occurs between researchers and program professionals. It is a process which I feel is indispensable in good research and evaluations. It *is* time consuming and expensive for both agency and researcher. It calls for rigorous scholarship on the part of the researcher both in his/her background work and field work. It calls for a real and extensive commitment out of program professionals. I suppose it is like milking a camel. It is difficult. It is painful. You will get kicked, spit on, and bruised. It takes a long time. People will think you crazy. But if you put your mind to it, really concentrate, and MEAN IT, REALLY MEAN IT, you will be able to milk a camel.

Measuring the Monetary Value of Lifesaving Programs

Jan Paul Acton
Economist
The Rand Corporation
Santa Monica, California

It very often happens in evaluating some program or other intervention that the issues ultimately boil down to a matter of economics. Specifically the question which must be answered is whether in view of its costs an intervention is worth doing. There are two distinct problems which have come to be known as cost-effectiveness and benefit-cost analysis. Jan Acton, in the two papers which follow, discusses these two types of analysis and tries to show how the more fundamental problem of benefit-cost ratio would be approached in the context of provision of emergency medical services.

I. Introduction

A multitude of public investment and regulatory decisions which have some effect on mortality and morbidity rates are made by legislatures, administrative agencies, and the courts every year. Typically, as in the case of highway safety engineering, the choice which confronts the public decision-maker is between reduced mortality rates and hence longer life expectancy for some group and more resources available for other purposes (e.g., additional miles of highway construction or a reduction in taxes). A decision to require something other than the minimum technologically feasible mortality rate reflects in effect a judgment that mortality (or safety) is not to be given lexical priority in public decisions over all other commodities which money can buy—a judgment which is certainly reasonable and in accord with everyday decisions made by households. If mortality is *not* to be given lexical priority, some other standard or procedure is needed to determine which projects are worthwhile. In particular, a procedure is needed for measuring the benefits of such programs in units which can be readily compared with the costs.[1]

In some constrained decision situations, the costs can be expressed in units of an identified commodity: For example, a school board may be faced with the decision of how much of its budget to spend on school bus safety, knowing that every additional dollar spent on bus monitors and drivers' salaries will reduce the quality of education by a certain amount. The choice between safety and the quality of education is easily understood and could be assessed directly according to the preferences of the public as represented by the school board. More generally, money allocated to safety will be taken from a fungible source which has many alternative uses. In such cases, there is no good alternative to measuring the cost of safety in dollar items, so that the evaluation of such a program will require the decision-maker to place a dollar value on safety, at least in an implicit sense. (Even in the school bus safety example, it is not appropriate to phrase the safety evaluation question in terms of educational quality units if changing school taxes is a viable option.)

How are we to go about placing a dollar value on the health and safety effects of a public program? The method which is in accord with the theoretical postulates of welfare economics is to measure benefit as the sum of all affected individuals' willingness to pay for the proposed program.[2] We can imagine each household being informed of the potential effect of the proposed program on its members' own safety and the safety of all those they care about, and then sending a ballot to the appropriate agency which indicates the maximum amount they would be willing to pay to have the program enacted. Their response will reflect the risk aversion, their anxiety of dying from the particular cause which is to be modified by the program, their financial circumstances, and the objective reduction in risk to them and their friends. If the aggregate willingness to pay exceeds the costs of the program, then the program is worthwhile in the sense that everyone could be made better off by its adoption: It is possible (though probably not administratively practicable) to charge each beneficiary less than it is worth to him and still cover the program costs. This "potential Pareto improvement" criterion is the formal theoretical justification for cost-benefit analysis, and it applies as well to evaluation of programs to reduce mortality or morbidity as to more traditional subjects like irrigation evaluation.[3]

This method, then would define the benefit of a program which can be expected to save ten "statistical" lives out of a population of 100,000 as the total value the 100,000 members of this popu-

lation place on having the probability of *each* individual's death reduced by one in 10,000. An alternative method, and the one which is actually used in almost all evaluations of public health and safety programs, is to attempt to actually place a money value on the lives that the program would be expected to save if it were adopted. In the example above, the "benefit" of the program would be 10V, where V represents the average "value of a human life." The method frequently used in practice for the heroic job of assessing V is to calculate the so-called "livelihood" measure[4]—the present value of lifetime earnings for a representative individual. The normative viewpoint which apparently motivates this approach is either that (1) people are properly thought of as the chattel of the state, and the loss of a life has a cost to the state comparable to the cost of a slave's death to his owner; or (2) the proper objective of public policy is to maximize Gross National Product.[5]

A third procedure for benefit valuation has not been employed in the past, but is potentially valuable. Since various public agencies and legislatures have been confronted with many decisions which in effect involve tradeoffs between dollars and mortality rates, there is considerable precedent for current decisions of a similar sort. Analyzing these precedents could help to increase the consistency of government decision-making.

Before proceeding to discuss these basic approaches to measuring the benefit of safety-enhancing programs in more detail, it is useful to indicate some of the seemingly related issues which, from a normative viewpoint, are in fact quite different. First, we are not dealing with the question of how much the government should spend to attempt to save the life of an identified individual (the coal miner trapped in a cave-in or the child in kidney failure) who is certain to die in the absence of government intervention. This is a very difficult issue because of, among other things, the symbolic importance of maintaining a public commitment to preserve life, which according to Calabresi and others is properly viewed differently from the safety investment issue.[6] Second, we are not attempting to determine the appropriate amount of compensation or punitive damages award (to either the individual or his survivors) for injury or death. While this issue is related to ours, in that court settlements in such cases may well influence the amount which private firms and households invest in safety, the relationship is complicated by equity considerations and a number of other considerations—including the desire to establish correct incentives for people whose actions influence mortality rates.[7] Third, we are not attempting to analyze the demand for life insurance, since this is determined by an individual's bequest motive and not by the value he places on his own safety.[8]

The remainder of the paper considers each of the procedures for benefit valuation mentioned above, but in reverse order. A final section summarizes the principle arguments and makes several recommendations for policy analysts.

II. Political Precedent

The logical first place to look for a source of standards for evaluating public programs which enhance health or safety is to the political process. If decisions regarding these programs tend to reflect a consistent set of values, then these values have a claim to political legitimacy and should be brought to light.

First, what does it mean for these decisions to be internally consistent? Investment and regulatory proposals differ in many dimensions, including the identity of the target population, the cause of death or disability which is to be curtailed, the nature and magnitude of the projected effect,[9] various side effects, and cost. To focus on the implicit valuations which such decisions place on improved mortality rates, two assumptions are useful: (1) Linearity: A program which reduces the probability of death by two in 1000 for each member of a specified group is worth twice as much as a program which causes only a one in 1000 reduction; and (2) Indifference to cause: The particular source of death which is to be curtailed by a program does not influence the program's value—all that counts is the number and perhaps characteristics of lives saved. If these assumptions are accepted, then a consistent procedure for assessing the benefit of programs is to value each of them by the number of lives which it is predicted will be saved, multiplied by some number representing what is often called the average "value of life" for the program's target population.[10] Precedent decisions can be analyzed to ascertain whether they reflect a consistently applied set of life values.

For any number of reasons it comes as no surprise that public program choices do not reflect the type of consistency defined above. One study which examined a number of lifesaving programs found implicit values of life which ranged from a few thousand dollars (in highway safety design) to over a million dollars (in an ejection system for an air force bomber).[11] To some extent this variability may reflect deviations from one or both of the simplifying assumptions stated above. For example, a higher and more expensive standard of safety for airplanes vis-a-vis highways may be justified by the argument that the threat of a crash seems to produce greater anxiety in air passengers than in auto passengers, even though the objective probabilities of death/mile are lower for the former group—this may generate a disproportionate demand for air safety. (In this vein one could also point to the disproportionate concern about

death by shark bite or being murdered by a stranger.)

Inevitably, however, much of the variability is the result of decentralized and varied decision-making processes, special political interests, and ignorance. Analyzing past decisions for precedents in defining the appropriate value of safety and health programs would be useful to the extent that it helped dispel this ignorance and yield understanding of the implications of consistency for decisions concerning programs under current consideration.

Ultimately, the study of precedent decisions does not yield an absolute standard by which to measure benefits of potential programs—it does offer a contingent standard which may be useful. If established program X is generally recognized as worthwhile, the proposed program Y offers a comparable increase in life expectancy/dollar expended, then there is a good argument for adopting program Y. In the absence of a consistent set of values generated by the political decision process, however, there remains a pressing need for benefit values calculated on the basis of more fundamental normative considerations. It is this need which, rightly or otherwise, is currently being filled by the "livelihood" procedure for life valuation.

III. Livelihood-Saving Measures of Value

Livelihood-saving is the most commonly used formal method for assessing the value of reducing mortality, and has been used as such for over 50 years.[12] This measure is based on the net present value of changes in the person's earnings stream.[13] By this criterion, if the expected livelihood-savings associated with a project exceed the costs of the project, it is worth undertaking,[14] otherwise the project is not worthwhile. Despite considerable discussion and use of livelihood-saving measures in the literature, there does not appear a clear statement of why it might be desirable to employ such a criterion for funding public programs. In particular, there is no reason to believe a priori that changes in earnings streams bear any direct relationship to what society values in health or safety program outputs.[15]

The livelihood-saving approach may have received the attention it has because it is relatively easy to apply and gives the impression of providing an unambiguous numerical answer. It is easy because the analyst can consult a table to determine the livelihood at different ages, identified by sex, race, and education.[16] The impression of numerical precision is more apparent than real, however. A number of important assumptions underlie the tables, and unless the decision-maker is conscious of their meaning, he may be unconsciously supporting a social judgment that he would reject if he faced it explicitly.

A. Intrinsic Shortcomings of Livelihood Approaches

The major objection to a livelihood evaluation is that it lacks a satisfactory normative justification. It is possible to infer from the way this approach is discussed in the literature that it is supposed to be justified by analogy to the economic procedure for valuing a machine or other piece of capital equipment. If a machine is accidentally destroyed, the resulting economic loss is equal to either (1) the cost of replacing the machine, or (2) the present value of the services which the machine *would* have provided if it had been saved—whichever is less. If the market for such machines is competitive, then measures (1) and (2) are equal, and both valid. Furthermore, the value of the machines' services is equal to the implicit or explicit rental price of the machine. People can be viewed as embodying "human capital," the services of which are rented in the labor market or used in home "production" (housecleaning, child care, etc.) The rental rate (wage rate) for labor services will under some assumptions reflect the value of such services in production. *If* we are to accept the notion that the social value of a life is equal to the value of the labor services the person provides, then the present value of the person's expected earnings (including "implicit" earnings from home production) is the appropriate measure of this value.

People are not machines, however. If we accept the view that production is not an end in itself for people, but rather a necessary intermediate step which allows us to enjoy the fruits of production, then the "human capital" approach is clearly inappropriate. Increases in safety and life expectancy help to ensure the continuation of an individual's ability to enjoy the pleasures of his life and the pleasure which his family and friends derive from a continuation of their relationship with him, and it is the value of prolonging this enjoyment which should be assessed in measuring the benefit of public programs which affect safety. While this hedonistic view would not be appropriate in a slave society (at least from the owner's viewpoint) or in a society dedicated solely to increasing the Gross National Product, it seems entirely appropriate in an individualistic society where the government is viewed as serving the public rather than vice versa.[17]

The livelihood procedure might still be accepted in practice if it could be demonstrated that it provides a reasonable approximation to a measure which does have conceptual validity—or even to our intuitive notions of what equitable policy requires. For some judgments at least, this type of justification is clearly lacking. For example, it is an inescapable conclusion of this criterion that society should spend no money on programs that extend the lives of fatally ill children because the programs would produce no change in their future

earnings. Furthermore, most persons would not agree that it is as important to save one worker earning $10,000 per year as it is to save two workers with similar personal and family characteristics, but each earning $5,000 per year. It is even more doubtful that most decision-makers would want to save men and women in proportions that depend on their earnings—even if a homemaker's services are valued at the wages of a domestic worker rather than at zero. For instance, the livelihood-saving calculation presented below shows that a white man in his 50's is valued more highly than a white woman in her 20's. If we were using livelihood-saving as the measure of value for government health programs, this means we would rather approve programs that save 55-year-old men than programs saving the same number of 25-year-old women. It also indicates that it is worth about twice as much to save one 25-year-old man as to save one 25-year-old woman.

It is doubtful that these magnitudes reflect the rate at which most people would want public lifesaving and morbidity-saving resources allocated. There is little direct evidence on this point about societal preferences, but what exists explicitly contradicts this implication of the livelihood approach. In Acton,[18] 91 persons were asked hypothetical questions about which person they would like to see saved if two seriously injured men arrived at any emergency ward and there were resources available to save only one of them.[19] The respondents had to choose between several different pairs of ages. Approximately one-third (31) of the respondents always chose to save the younger person; 39 expressed a preference that was single-peaked in age (peaks generally occurred between 20 and 30 years of age as does the human capital curve); and 8 were indifferent to all age pairs. (The remainder were multipeaked or inconsistent rankings.) Thus, somewhat less than half the respondents expressed a desire to save lives identified by age that corresponds to the shape of the livelihood curve.

The livelihood measure assigns a higher value to men than to women at almost all ages, but this sample rejected such a ranking when asked to select a man or a woman of identified ages in the emergency-ward question above. The majority of persons (53) selected only on the basis of age and matched the same ranking they had expressed when selecting between two men. Nine respondents always selected the man over the woman, and nine always selected the woman over the man. In one question, the respondents were asked to choose between a 30-year-old man and a 30-year-old woman. Thirty-seven chose the man, 43 chose the woman, and 11 expressed indifference.

We are not aware of any other systematic empirical evidence about people's preferences for saving lives identified by age or by sex. However, this empirical evidence, along with casual observation of attitudes for public programs, suggests that a majority of people would at least reject the relative value of saving men and women that is implied by the simple livelihood method. In the provision of public services, where objectives may include allowance for factors such as income redistribution, and externalities such as the numbers of dependents that will be orphaned, the social evaluation may even vary *inversely* within measures of livelihood involved!

Even if we were satisfied that the livelihood procedure formed a conceptually sound basis for public program evaluation, an important practical issue remains to be resolved; Market earnings in some cases do not equal the productivity of an individual's labor.

B. The Issue of Earnings vs. Productivity

A person's earnings may differ significantly from his productivity for a number of reasons. For instance, workers in a strong union may earn considerably more than workers doing identical nonunionized work. Some groups may face earnings discrimination because of their race, ethnicity, or sex. Some people (e.g., people with job seniority) may be receiving an income substantially above their productivity. The livelihood measure is blind to these distortions. It merely says to add up the earnings of people who may be affected by different programs, and select the ones that save the most earnings. Since diseases typically do not affect different racial, sexual, or socioeconomic groups uniformly, a criterion that depends on earning differences among these groups will necessarily slant public programs in particular directions. If some diseases are found more often in people with higher earnings, the rule says to devote your attention and resources to these diseases.

The undesirable nature of this criterion is brought home acutely when we consider the implications for the treatment of women (although it applies in less extreme form to any case where wages do not reflect productivity). The national product accounts do not include the homemaker services of women if they are not purchased; but to include them from a measure of project benefit will seriously undervalue programs that affect women. The most common procedure is to value homemaker services at the full-time earnings of a domestic worker; compare Weisbrod,[20] Klarman,[21] and Rice.[22] Various arbitrary weighting rules have also been used (see, for example, Feldstein [23]).

Using the earnings of a domestic servant is only partly satisfactory, however. In the first place, the homemaker may be providing quantity or quality of services that are not available in the market. For instance, when we observe a woman with advanced education who could take a job paying two or three times a domestic servant's income, she may be staying home to raise her small child be-

cause she feels the first few years are important and because she does not feel she could hire such high-quality nurturing for her child. Under the circumstances, using the domestic servant's earnings will understate the value of this woman's home activities, as she sees them. In such circumstances, we could argue that her services at home should be valued at least as highly as the highest salary the woman could earn.[24] However, we probably do not really wish to adopt the implications of such reasoning. After all, many people accept jobs at a salary less than the maximum they could command in the market. They may do this in order to have better working conditions or in order to pursue a particular type of work. In the extreme, the implication of this foregone opportunity argument is that we should value everyone's services—men's and women's—at the highest possible wage they could earn. Ignoring the readjustment this would cause in the general wage scale, such a recalculation would raise the implicit earnings of society considerably.

A second objection to the standard treatment of home production is that it is assymetric with respect to sex. After all, women are not the only workers around the home. Morgan et al.[25] and Walker and Gauger [26] surveyed people about the hours they spend working around the house. They found that men spend between about one-eighth and one-third as much time as do women, depending on the employment status of the woman, and the ages and family sizes involved.[27] If we are imputing a value to individuals for their home production, then it seems appropriate to add an element to the man's livelihood calculation.

The third objection to the standard treatment of home production lies in the treatment of older women, especially over 65 years of age. Rice and Cooper [28] attributed a full domestic worker's income to nonemployed women over 65, causing their livelihood to exceed significantly that of a man over 65. One could speculate that women over 65 start to slow down in their household activities, but it is difficult to find data. Walker and Gauger [29] did not survey older women. We aanalyzed the results of the Productive Americans Survey (partially reported in Morgan et al.[30]). The number of observations is relatively small in the over-65 age group, but there appears to be a downturn in average number of hours worked at home by women and an increase in the hours worked by men. Women's hours declined about 19 percent in the over-65 age group and men's hours increased about 17 percent. This leaves women over 65 reporting about 35 hours of housework per week and men reporting about 6½ hours. These figures may represent an overstatement of true contribution if productivity falls significantly in this age group. Furthermore, there may be some reporting error if the respondents have little else to do and therefore claim that most of their times goes to housekeeping.

Since there are no compelling theoretical arguments for one rule over another in accounting for household production, livelihood tables can be generated under a variety of assumptions about the value of women's and men's contributions.[31] These calculations show significant variation in the livelihood, especially in the upper ranges, depending on the assumptions employed. For illustrations, Figs. 1 and 2 plot the livelihood at different ages for a four-way breakdown of sex and race under two of the assumptions possible for treating home production. The assumptions behind the calculations are discussed in more detail in Acton,[32] but briefly, Fig. 1 (Assumption 1–1) assigns a value of $4800 for the domestic work of nonworking women.[33] Figure 2 (Assumption 3–3) assigns a variable amount to women's homemaker function (depending on their employment status) and a uniform amount to men. After 64 years of age, women's contribution is reduced (19 percent) to reflect a drop in household activities, and men's is increased (17 percent). A 4 percent net discount rate is used for both figures.

We do not intend to focus on the nature of livelihood at different points in life or to concentrate on differences among races and sexes (although they are already quite substantial). These plots, however, serve to emphasize the substantial variability due to alternative assumptions about the valuation of household activities and the substantial impact this has on the relative and absolute amount assigned to women by this criterion. The effect of these alternative assumptions is significant at all ages—but it is especially noteworthy in the over-65 age range where a substantial amount of mortality and morbidity is involved from such prominent ailments as heart and circulatory diseases and cancer.

The plots in Figs. 1 and 2 show a close similarity between the livelihood for white females (WF) and all other females (AOF). This is due to the relatively low work rates of women, combined with the assumption that all nonworking women are assigned the same value of household services regardless of race. The differences between white males (WM) and all other males (AOM) is about the same under the two assumptions and measures about $60,000 higher for white men in their late 20's than nonwhite men in the same age. The difference between sexes is dramatic—with the livelihood of white males at its peak about 2½ times the level of white females at its peak under Assumption 1–1. When the household production of working men and women is given an imputed value (Assumption 3–3), the differences between the sexes narrow considerably. At its peak, white men's livelihood is only 1.7 times that of white women. The male:female ratio is even closer for nonwhites.

Fig. 1—Basic human capital, assumption 1–1, interest rate = 0.04

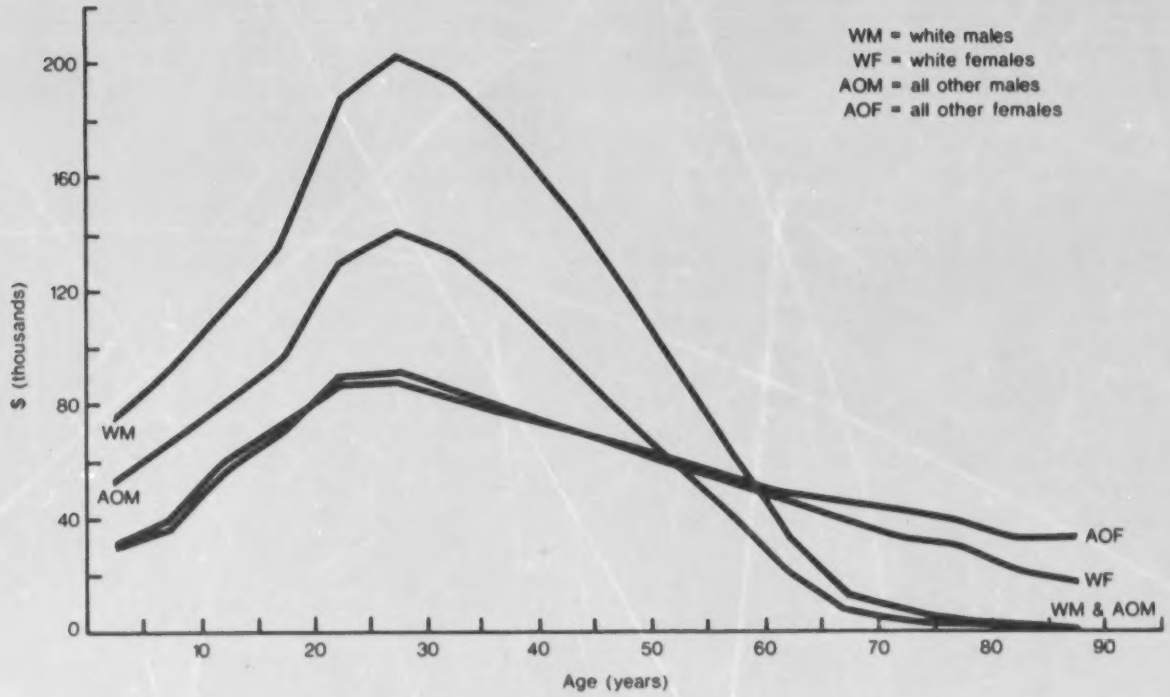
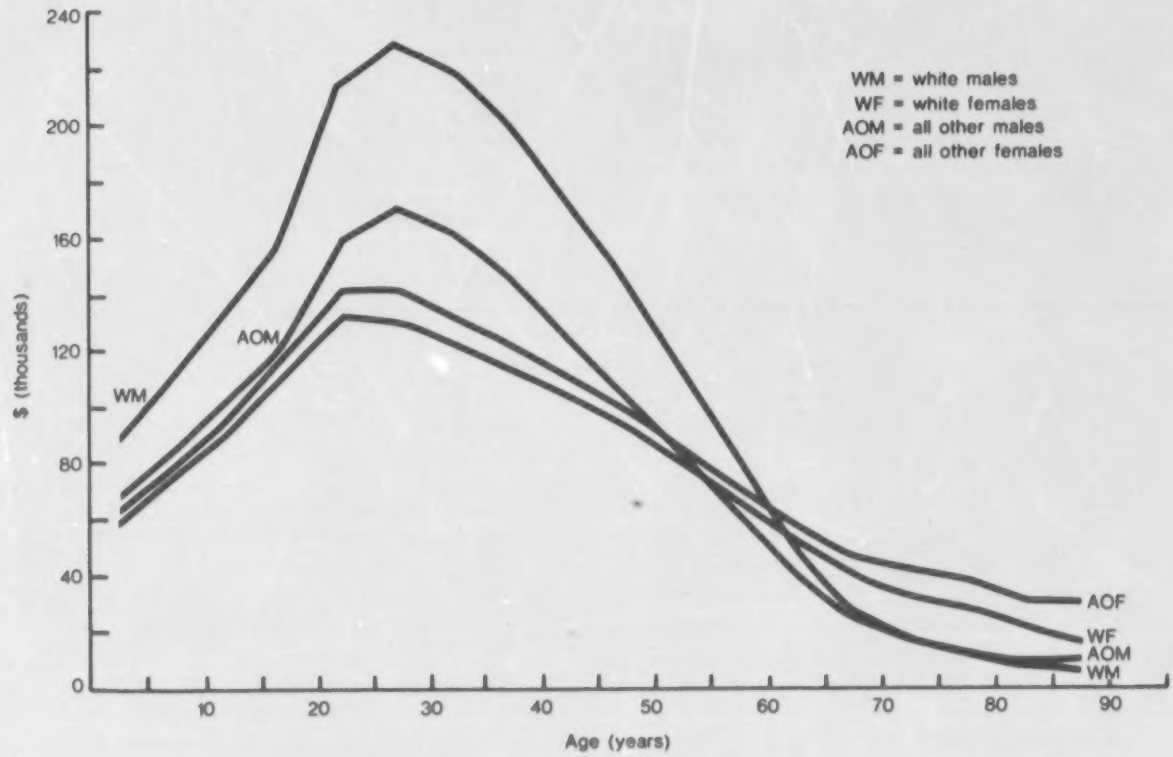


Fig. 2—Basic human capital, assumption 3–3, interest rate = 0.04



The other major effect of the different assumptions comes in the crossover between male and female livelihood in the upper age brackets. Under Assumption 1–1, female livelihood crosses male between 50 and 60 years of age—due both to the lower life expectancy of men and the fact that women are assigned a value of household production while the generally retired men are not. Consequently, over 65 years of age, male livelihood falls to extremely low levels, while female livelihood remains between $20,000 and $40,000. Under Assumption 3–3, when a greater value is assigned to household production for men and for working women, the reversal for white men's and women's livelihood is postponed to the early 60's, and the livelihood of men is higher than before in both relative and absolute terms. The reversal for nonwhites is pushed to a lower age, but the difference at all ages is narrowed considerably.

IV. The Willingness-to-Pay Measure of Value

A fundamental assumption of the willingness-to-pay procedure is that individual's preferences should count—that citizens can and should play a role in policymaking for governmental services that affect them directly. Their health, their friends, their taxes, their pain and suffering, and their welfare are at stake. Understandably, they have an interest in the public activities that may be undertaken. Individuals are the ultimate recipients of the impact of programs.

Political justifications for using individual preferences go back at least to the 17th century and include the desire for no taxation without representation. Economic arguments for using individual's preferences date to the 19th century and include the utilitarian principles of Bentham. Dupuit,[34] a French engineer, argued that the nature and amount of public transportation facilities should be determined by what the potential users would be willing to pay for using it. Most contemporary economists who study public policy evaluation agree that an approach based on individual values is correct in principle.[35]

The "potential Pareto improvement" standard which justifies the willingness-to-pay procedure has been criticized because it makes the estimated dollar benefit of a program dependent on the income distribution. This dependence has been criticized either because (1) it is felt that the income distribution is inequitable and hence not a just basis of public program evaluation; or (2) it is felt that *whether or not* the income distribution is equitable it is simply not an appropriate basis for determining the production and distribution of certain goods (possibly including adequate health care and safety) which are, like the vote, properly considered noncontingent privileges of membership in society.[36] The problem which has not been solved by critics is to devise an alternative benefit measure which satisfies such objections. The livelihood measure is even more directly tied to income distribution (viz., by definition) than is the willingness-to-pay measure, and it is not impossible that precendent political decisions were influenced by the economic power of various interest groups.

The principle practical problems with the willingness-to-pay procedure for benefit estimation is that developing accurate assessments of individuals' willingness-to-pay is difficult and expensive, and the validity of published attempts to apply various estimation techniques is questionable. Furthermore, the extent to which estimates of a particular population group's willingness to pay for a particular safety-enhancing project can be applied to other groups and other types of projects is unknown.

The two principle methods for measuring the values a household would place on a prospective public project are (1) Inferences of how much the household values mortality reduction based on observations of the implicit value the household places on safety and health in making private consumption and job-selection decisions; and (2) Survey questionnaires which ask household heads to state their willingness-to-pay for the program benefit which is under consideration.

A. Implicit Values

We can, in principle, infer the values individuals attach to mortality- and morbidity-reduction in the same manner as was proposed for governmental actions (Section II above). Such a revealed preference approach is followed with most market-produced goods that have few externalities.[37] We need not go into a detailed survey of relative preferences for, say, apples and oranges. People reveal the preferences they attach by their market behavior. This is the method we would like to use if we want to measure individuals' true preferences for the programs. It presents the strongest claim to validity because the people have to back up their preferences with action, and they do it in the context of other everyday decisions for spending money.[38] These choices may include the purchase of safety devices (for example, seat belts), a marginal expenditure on health items (perhaps a doctor's examination and some antibiotics for an infection), or the premium demanded for accepting an elevated risk (for instance, higher wages for extrahazardous employment).

Recent studies by Thaler,[39] Thaler and Rosen,[40] Smith,[41] and Usher[42] have provided measures of implicit willingness to pay for lifesaving. Thaler, Thaler and Rosen, and Smith examine the higher wages paid in occupations with above-average risk of death for evidence about the implicit value of lifesaving. Usher employs a life-cycle model of utility maximization and infers the trade-off between consumption and probability of survival from a time series of the national income

accounts and mortality statistics. Both approaches have the potential of overcoming some reservations about the survey-based willingness-to-pay approach because they examine behavior revealed through market activity and therefore have stronger claims to validity and stability than existent survey results.

Since the two Thaler studies and the Smith study rest on market wages, they have some drawbacks in common with the livelihood-saving approach. First, the measure requires that the person be working to determine a value. Therefore, it is difficult to determine the appropriate value for housewives, children, retired persons, and others who are not paid for their work. A second criticism relates to the representativeness of this group observed in riskier occupations. Presumably, those who are least risk-averse will enter a given occupation before those who are more risk-averse, all other things the same. Consequently, lower risk premiums will be paid to those who select the occupation that would be necessary to compensate a randomly chosen individual who was subjected to that level of risk, and these measures will be a lower bound on "society's value." Third, the extra pay is compensation for assuming an *above*-average risk, and for that reason may not provide an appropriate measure of value for programs which are designed to reduce risk. The compensation which a risk-averse person would require to accept a Δ p increase in the probability of his own death is greater than the amount he would be willing to pay for a Δ p reduction in this probability—although the amounts will be close to one another for small Δ p. Fourth, the wage-premium observed will not necessarily reflect the externalities (to family and/or society) associated with a person's death—although the externalities will be better captured with this measure than with the livelihood-saving approach if the employee includes his family in the job-choice decision and requires that the wage-differential be adequate to compensate them for his increased risk as well. Fifth, it is difficult to identify what portions of differences in compensation are due to the additional risk of death, risk of injury, and other working conditions. Sixth, although it is not a general phenomenon, there may be some occupations in which the participants receive some utility from the risk, and therefore the compensation is inadequate for a normal person. Being a stock car racer or being a test pilot may be extreme examples, but this consideration may be reflected to some degree in a number of occupations, some of which are included in Thaler's calculations. Finally, at the conceptual level, we do not know for certain what risks of death or injury the individual assumed were in force when he accepted the wage offer. Given the difficulty Thaler seems to have had in getting good data on death rates by occupation, the amount of uncertainty a given individual faces about the risk at a particular job site may be substantial.

On the empirical side, Thaler found significant variation in implicit valuation depending on the data source used. With one data file, he inferred a value of between $176,000 and $260,000 per expected life saved (for a reduction in probability of 0.001 per year), which is remarkably close to the peak human capital value observed for young men and to the explicit willingness of pay obtained by Acton [43] in his survey for a reduction of 0.001 in heart attack death rate. On the other hand, the value implicit in the Bureau of Labor Statistics injury data was over $2.6 million per expected life. Furthermore, Thaler's regression results with the BLS data yield an incorrect sign for the coefficient of risk of injury. The regression with the first data file did not include a variable for risk of injury, so his results are subject to omitted variable bias, and the difference between the first and second estimates were even more extreme than they appear.[44]

Usher's study is an imaginative use of the (Canadian) national income accounts to infer a tradeoff between consumption over a life cycle and resources devoted to death reduction. He makes utility solely a function of consumption in each time period (which is equal in all time periods) as well as the probability of surviving, and employs strong assumptions about the form of the utility function to make his estimates. Given the strong assumption about functional form, the potentially severe aggregation bias from using such highly aggregated data to infer a utility function for individuals, and the absence of an indication of the level of statistical significance, we may wish to place most emphasis on the qualitative findings. Usher's model implies that the value per expected life saved is greatest at a very young age (it peaks around age 2 for plausible values of his parameters) and decreases through increasing age. Its value in the age sample 20–30 is very similar to the human capital values reported for white males by Rice and Cooper.[45] Since utility is a function solely of consumption (not earnings) and since he assumes that every one consumes the same amount in each year of life, there is no difference between the value assigned to men and women in his model.

B. Explicit Statements of Individuals

The survey approach [46] permits measurement of the entity which is directly appropriate to evaluating a proposed public project—the maximum amount each affected household would be willing to pay to have the project adopted. In theory this procedure requires no assumptions about individual preferences (e.g., linearity, indifference to cause, absence of externalities) which other techniques require. Since the expense of conducting a special survey for every proposed project would be prohibitive, however, in practice

we would want to generalize from the results of one survey in order to assess other project proposals—such generalizations will of course require some assumptions on preferences.[47]

While willingness-to-pay surveys have been conducted successfully in recreation program evaluation,[48] the only published survey we have found of willingness to pay for health programs is contained in Acton,[49] and that survey deals only with programs that reduce chances of sudden accidental death or heart attack death. It sought preliminary evidence on the feasibility of applying willingness-to-pay responses to actual program evaluation and addressed several questions:

- Can questions be formulated that in principle get at willingness to pay?
- Do people seem willing to answer and are they relatively comfortable in answering such questions?
- Are the responses people make subject to a rational interpretation?
- What seem to be the major factors influencing stated willingness to pay?

In total, approximately 125 persons were questioned about their willingness to pay for heart attack mortality reduction.[50] People were posed four types of questions:

1. Age choice questions—Which of two seriously injured would you like to see saved in an emergency? Those results were discussed above in the critique of livelihood-saving measures.
2. Live in the community—How much would you be willing to pay to have a heart attack ambulance that is expected to save X lives per year of the 10,000 people living around you?
3. Advice willingness to pay—Suppose your neighbor has just been told his risk of heart attack is Y per year, and his chances of dying if he has a heart attack are Z. How much do you think he should be willing to pay per year for a heart attack program that would reduce his chances of dying to Z*?
4. Own willingness to pay—Suppose your doctor tells you your chances of a heart attack are Y per year, and your chances of death, given the heart attack, are Z. How much are you willing to pay per year for a heart attack program that can reduce your chances of dying to Z*?

Each respondent answered 26 questions of type (1), two questions of type (2), and four questions each of types (3) and (4).

The results showed that we can pose questions that get at the underlying issues of willingness to pay. Furthermore, people were willing to complete the interview and seemed relatively comfortable and responsive in doing so (the refusal and breakoff rates were negligible). The question of rational interpretation of the responses was not clearly resolved in a single survey of this size. Responses varied significantly from one individual to the next (only part of this could be explained as sampling variance due to sample size). High variation per se is neither unexpected nor undesirable for these types of questions. We expect preferences and attitudes to vary from one individual to the next, even for identical expected benefits offered to individuals who appear to be similar in the socioeconomic and demographic profiles. Nevertheless, the responses of most persons could be given a rational interpretation, and predicted effects were found for important explanatory variables such as income, wealth, age, and sex. The empirical results are discussed in detail in Acton.[51] Briefly, the principal statistically significant findings were that willingness-to-pay responses increase with increasing probability of death and with greater reductions that are offered—but not in a linear fashion.[52] Second, willingness-to-pay responses are greater the more concretely and immediately the hypothetical program is related to the individual.[53]

If such willingness-to-pay responses were to be used routinely for program evaluation, we would wish to conduct a survey of a greater number of respondents (appropriately selected for statistical representativeness) where the questions included several different probabilities of mortality, morbidity, and several different reductions in the values of each health consequence. If it appeared conceptually or empirically desirable, separate sets of questions for major categories of diseases or risks should be prepared (for instance, heart diseases, cancer, accidents, and so forth). If satisfactory, statistically significant willingness-to-pay relationships were found, then it would probably be most efficient to use the results to multivariate regression equations to estimate the aggregate willingness to pay associated with a particular program—taking account of the socioeconomic and demographic characteristics of the population affected and the anticipated changes in probabilities.

A number of issues are still left open in the feasibility of a survey-based method for eliciting value. These include the validity of the responses, their stability and replicability, problems with understanding and processing the information in these hypothetical situations, and strategic behavior in responding.

The validity of responses to willingness-to-pay questions has not been examined empirically. Indeed, it is not clear that the validity can ever be firmly established. A rigorous test of validity might be to survey a group of people and then come back and actually market the goods that had been described (say a heart attack ambulance) or raise their taxes in accordance with responses. Some people might refuse to act in accordance with their

previous responses because of intervening factors which may be difficult to control for and which the respondent cannot even articulate.[54]

The stability and replicability of these preliminary results have not been demonstrated. Further empirical work is clearly needed to see if the same people respond with a reasonable stable set of preferences when resurveyed at a later date. Furthermore, we should see if the results can be replicated in other geographic areas with different socioeconomic and ethnic samples.

We face several competing objectives in asking questions that are both realistic and yet understandable for the respondents. Since many of the situations we pose to people are hypothetical (either the disease state or the consequences of the programs), we are uncertain about the individual's comprehension of the situation. For instance, although heart disease accounts for about ⅓ of all deaths per year, the realistic chance a person has of dying from a heart attack is less than 1 per 100 per year for the majority of adults. We are, as yet, uncertain about how well people understand and process such numbers.

Similarly, we do not necessarily know how well people understand the nature of certain disability states or recoveries. The operationally relevant point, however, is whether they understand the situation well enough during an interview that their preferences do not change significantly if a decision is made to inaugurate the program. The most direct way to test this assumption is to examine the stability of responses over time.

A fourth unresolved issue in willingness-to-pay elicitation is whether people will engage in strategic behavior when they respond. Lindahl [55] observed that when you try to find out people's preferences for public programs, they may have an incentive to underrepresent their true valuation if their taxes depend on their stated value. Acton [56] and Bohm [57] observed that the opposite case may also exist if people think the decision whether or not to have the program is based on aggregate value, but the cost-sharing rule is determined by a different rule. Under these circumstances, if the person feels he will be called on to bear a small proportion of the costs for a project he wants, he should overrepresent his willingness to pay for it. Dreze and Poussin [58] have shown that under some circumstances, people will have the correct incentives to reveal their true preferences for public goods that are already being produced. Bohm [59] suggests that people be posed questions where the payment rule is deliberately specified as yet-to-be determined. In this manner, he expects to cancel the incentives to over- or underrepresent true feelings, because people will not be able to select a strategy for a misrepresentation of references that is guaranteed to make them better off than telling the truth.

Bohm [60] conducted an experiment to see how sensitive willingness-to-pay responses were to question wording and to analyze whether strategic behavior seemed present. The sample does not purport to be fully representative (only 211 of 605 randomly selected residents of Stockholm agreed to participate), but the experimental design is intriguing and to the point. He paid the volunteers Kr.50 ($10) for a one-hour "interview" about television programs. When the respondents came to the studio, they were told the interview was delayed and they were put in a room with TV screens and given an opportunity to watch a comedy show with two very popular comedians. They were given the impression that several other respondents were in similar rooms around the building and that the program would be shown only if the aggregate willingness to pay exceeded the cost associated (Kr.500). The different respondents were randomly given different instructions about what the decision rule for actual showing would be.[61] If people were behaving strategically, some instructions should cause significantly higher responses than other instructions. Bohm's empirical results show no statistically significant difference (at 5 percent) in the responses from one question form to another.

At the moment, we can conclude that although strategic misrepresentation may exist in principle in the willingness-to-pay context, it has not been demonstrated to be a significant empirical factor. At the pragmatic level, it is relatively unlikely to be a serious problem with preliminary efforts to assess people's values, because people are not accustomed to having their tax bill react to such statements of value.[62]

Many of these potential problems in implementing a willingness-to-pay measure will be clarified only with additional empirical evidence. For instance, the estimates of the true variance of responses in society and the mean value of the responses can only be judged by conducting surveys of representative populations of respondents. Similarly, the reproducibility and stability of responses over time can be measured, but have not yet been explored empirically. Some of the more basic concerns about the validity of the responses and the internal consistency of a given person's responses are more difficult to resolve. We have crude measures of what "internal consistency" means, but to demonstrate rigorously its existence (or nonexistence) hard thinking is needed. An interactive process of both conceptual development and refined empirical evidence seems to be the most viable strategy for furthering our understanding in both areas. Furthermore, if done with some foreplanning, we can also provide useful

interim survey results that can be used as one measure of social impact valuation for current evaluation efforts.

V. Conclusion

There are important *conceptual* and *empirical* differences between approaches to evaluation reviewed here. The choice of method is important and may change the ranking and value of health or safety programs significantly. The selection of a particular method involves tradeoffs between ease of application and conceptual soundness. The livelihood-saving approach is easy to apply (and has been used frequently in the past), but it has a number of drawbacks when its implications are examined in detail. An approach based on individual preferences (operationally, what people are willing to pay) meets the drawbacks of the livelihood approach and is conceptually most satisfactory. Preliminary evidence suggests that it is feasible to ask for explicit statements and that meaningful answers result, but a number of problems may arise in implementation on a large scale. There has been very little empirical experience with measuring implicit value or with conducting surveys of people's willingness to pay for public programs. In the revealed preference approaches we may not observe a representative group of people, and it may be difficult to know with certainty that observed behavioral differences should be attributed only to differences in level of risk. Correspondingly, we do not know what the stability of survey responses is over time nor what the sample variance is likely to be. Furthermore, the validity and internal consistency of these responses is not yet established. It is difficult to specify rigorous tests of the external validity of these sorts of questions, but an interactive development of the conceptual underpinnings and empirical evidence provides promise of sharpening our understanding.

For many actual evaluations, both the livelihood-saving approach (with its known drawbacks) and an imperfect, crudely measured, willingness-to-pay methodology are clearly superior to no formal analysis. First, the analysis is frequently an order-of-magnitude evaluation. Under these circumstances, the drawbacks or questions we have about either approach are second-order magnitudes and do not affect the conclusion whether or not to undertake the program. Second, employing both criteria to see if they yield the same conclusion can reinforce one's confidence in the robustness of the decision. Third, in the range of expected effectiveness for many realistic programs, the approaches frequently lead to reasonably close measures and value.[63]

When given a choice between livelihood-saving or willingness to pay as a basis for evaluating social impact, a strong case can be made for the conceptual superiority of willingness to pay. The livelihood measure does not bear any necessary relationship to what people want in the way of public programs. If we decide to fund programs by this criterion, we know that we could, in general, raise adequate revenues by taxing those whose livelihood is extended.[64] However, this criterion does not guarantee that society or any individual is made better off by adopting the program.

An individual preference approach (based on willingness to pay) does provide us with an assurance that society is made better off in some sense by the programs that pass the criterion. By approving only programs such that people are willing to pay, in the aggregate, more than the programs cost, we can make a strong case that society as a whole gains. It is clear that in general the program will be funded in a manner such that some people gain and some lose with a particular implementation. Nevertheless, since the aggregate *willingness* to pay exceeds the cost, it would be possible to spread the costs such that no one was made worse off by the program. That is, with the criterion we identify potential Pareto superior moves for society. Every member can be at least as well off as he was without the program, and at least one person is better off.

Although we started this paper with the objective of identifying means of placing a value on reductions in probability of death or disability, we should recognize that it may not be possible (or desirable) to have a unique value that can be used in several different contexts. Instead, it may turn out that preferences are such that we have one value for a change in probability for cancer death, another value for a change in probability of heart attack death, and yet a third value for change in probability of accidental death—even for similar persons and identical starting risks and reduction in risks. Given the diversity of values now implicit in public decisionmaking, such a finding would not be unexpected. Furthermore, analysts like Zeckhauser[65] argue that the process by which public decisions are made may be at least as important as the actual numerical values used. An appropriate strategy for the decisionmaker charged with evaluating lifesaving programs before additional methodological and empirical research takes place may be to employ more than one of the techniques discussed. When the different approaches yield similar conclusions, he can gain confidence from the fact that his evaluation does not seem to be sensitive to the values employed. When they yield sharply different conclusions, he can probe his own preferences or seek additional evidence about the willingness to pay of the target population.

Footnotes

* Economist, The Rand Corporation, Santa Monica, California. I wish to acknowledge with gratitude the comments of P. Cook, W. Manning, B. Mitchell, J. Newhouse, J. Vaupel, M. Weinstein, and A. Williams. The views are those of the author and do not necessarily reflect those of the Rand Corporation or any of its corporate sponsors.

1. Formal prospective evaluation of governmental programs, as discussed here, is a relatively young discipline. Water resource allocation has the longest history in the U.S., having been charged since the 1930's to determine "if the benefits to whomsoever they accrue are in excess of the costs." (From Flood Control Act of 1936, quoted in A.R. Prest and R. Turvey, "Cost Benefit Analysis: A Survey," in SURVEYS OF ECONOMIC THEORY, St. Martin's, New York, p. 150 (1966)). Most of these applications in water resources have been limited to economic benefits and costs, although considerations such as recreational values and their distribution have been added; see, for example, B. Weisbrod, "Income Redistribution Effects in Benefit Cost Analysis," in Stuart Chase (ed.), PROBLEMS IN PUBLIC EXPENDITURE ANALYSIS, The Brookings Institution, Washington, D.C., 177–209, (1968).

A number of economists have reviewed various aspects of the evaluation literature. Prest and Turvey (Id.) have a good background review of the cost-benefit literature. P. Steiner (PUBLIC EXPENDITURE BUDGETING, The Brookings Institution, Washington, D.C. (1969)) focuses on a number of issues in program budgeting for federal programs. H. Klarman reviews literature related to health evaluation, focusing on the evaluation of health technology in "Application of Cost-Benefit Analysis to Health Systems Technology," in Morris Cotten (ed.), TECHNOLOGY AND HEALTH CARE SYSTEMS IN THE 1980's, DHEW Publication No. HRA 74-3011, Washington, D.C. (1973), NTIS PB No. 220 613 266p. H.R. Thaler ("The Value of Saving a Life: A Market Estimate," Ph.D. dissertation, Department of Economics, University of Rochester, New York (1974)) reviews some historical attempts at valuation of lifesaving, and R. Zeckhauser ("Procedures for Valuing Lives," PUBLIC POLICY, Vol. 23, No. 4, 420–463 (Fall 1975)) provides a discussion of some recent applications. There are several essays on public expenditure in general. Dorfman and Chase have edited works focusing on particular problems of public expenditure evaluation; see R. Dorfman, MEASURING THE BENEFITS OF GOVERNMENT INVESTMENTS, The Brookings Institution, Washington, D.C., (1965), and S.B. Chase, PROBLEMS IN PUBLIC EXPENDITURE ANALYSIS, The Brookings Institution, Washington, D.C. (1968). R.H. Haveman and J. Margolis have edited a (sometimes revised) set of essays on the Planning, Programming, Budgeting System (PPBS) experience by a number of practitioners and critics, titled PUBLIC EXPENDITURES AND PUBLIC ANALYSIS, Markham, Chicago (1970). Some of the most extensive and successful applications of formal analysis have been in the defense area. Although they have tended to be cost-effective rather than cost-benefit analysis (i.e., How can we best achieve a defense or tactical or strategic posture without asking how expensive a posture we should have?), some techniques developed there from the basis of analysis, especially regarding the general structuring of decisionmaking under uncertainty and the quantification of uncertain outcomes. A good introduction to this systematic approach to analysis, with a description of a variety of techniques, is found in a collection of essays eidted by E.S. Quade and W.I. Boucher, SYSTEMS ANALYSIS IN POLICY PLANNING, American Elsevier, New York (1968).

2. See in general E.J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," JOURNAL OF POLITICAL ECONOMY, Vol. 79, No. 4, 687–705 (1971). An interesting discussion of whose interests should be reflected in benefit valuation which considers the intergenerational problems is to be found in J.A. Dowie, "Valuing the Benefits of Health Improvement," AUSTRALIAN ECONOMIC PAPERS, Vol. 9, No. 11, 93ff (1970).

3. This criterion was originally proposed by both N. Kaldor, "Welfare Propositions of Economics and Interpersonal Comparisons of Utility," ECONOMIC JOURNAL, Vol. 49 (1939); and J.R. Hicks, "The Foundations of Welfare Economics," ECONOMIC JOURNAL, Vol. 49 (1939). A good recent discussion in the "valuing lives" context is J. Hirshlerfer, "The Economic Approach to Risk-Benefit Analysis," in David Okrent (ed.) RISK-BENEFIT METHODOLOGY AND APPLICATIONS (processed) UCLA-ENG-7598 (December 1975).

4. A term due to Schelling (T. Schelling, "The Life You Save May Be Your Own," in S. Chase, ed., PROBLEMS IN PUBLIC EXPENDITURE ANALYSIS, The Brookings Institution, Washington, D.C., 127–176 (1968))—as distinct from the lifesaving, or willingness-to-pay, approach.

5. See Mishan, note 2 *supra*.

6. G. Calabresi, THE COST OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS, Yale Univ. Press, New Haven (1975).

7. R. Posner, (ECONOMIC ANALYSIS OF THE LAW, Little Brown, and Company, Boston (1972)) R.M. McKean ("Products Liability: Implications of Some Changing Property Rights," QUARTERLY JOURNAL OF ECONOMICS, Vol. LXXXIV, No. 4, 611–626 (Nov. 1970)) have explored conditions under which economic efficiency is improved by

assigning liability to one party (say the producer of a good) rather than permitting the market to supply (or fail to supply) products that provide reductions in risk. Although, in general, these liability solutions imposed to improve economic efficiency will understate the value of lifesaving or disability saving that would be inferred from a direct assessment or willingness to pay, they cannot be used as an unambiguous lower bound because of transactions costs and lack of perfect information, possible differences between the group determining the law and those engaged in the transaction, punitive elements to settlements, or differences between the group affected ex ante and the group being compensated ex post.

8. See R. Eisner and R. Strotz, "Flight Insurance and the Theory of Choice," JOURNAL OF POLITICAL ECONOMY, Vol. 69, No. 4, 356–368 (August 1961).

9. J.E. Cohen ("Livelihood Benefits of Small Improvements in the Life Table," HEALTH SERVICES RESEARCH, 82–96, (Spring 1975)) reminds us that it is crucial to make clear the time course of the benefit for epidemiological as well as valuational reasons. Frequently, analysts have in mind a program that offers a reduction in possibility of death that is effective for one year at a time. Cohen points out that some program benefits may be more accurately characterized in a different manner, and that the alternative definition may make a large difference in the measured benefit. He defines a "curative" benefit as one that offers a person a one-time save (or reduction in probability of death) from a disease, regardless of the age at which it occurs, and then the person falls back into the normal risk pool. He defines a "preventive" benefit as one that eliminates a particular cause of death entirely. Cohen shows that substantial differences can arise in the measured total benefit when a curative or preventive benefit rather than a one-year explosure benefit is involved. In the case of kidney disease for U.S. males, his calculations yield a total benefit about 22 times as large as that of J. Hallan, et al., THE ECONOMIC COST OF KIDNEY DISEASE AND RELATED DISEASES OF THE URINARY SYSTEM. PHS Pub. No. 1940, U.S.G.P.O., Washington, D.C. (1968).

10. It should be noted that while the "value of life" terminology is convenient and frequently encountered within the philosophical framework of the livelihood procedure, it is strictly accurate only because of the linearity assumption. If decision makers are non-linear with respect to livelihood saving (eg., if they are not indifferent between (a) saving one person's life [and livelihood] with *certainty* and (b) saving one hundredth each of 100 persons' livelihood), then one cannot even speak of *the* "value of a life" within the context of the livelihood measure. Within the context of willingness-to-pay measures, it is meaningless to speak of "the value of a life." In general, one can only refer to the expected value per life saved *at a given initial risk of death* and *for a given reduction in risk*. Suppose a given individual has an initial risk of death P, and is offered a chance to reduce it by Δ P. If he will be willing to pay an amount, X, to reduce the risk, then we may refer to the value Y (which equals X/ Δ P) as the expected value per life saved for this set of circumstances. (It can also be viewed as the amount that a large number of people similarly affected and with similar tastes would pay, on the average, for each life saved in their group.) In general (because of risk aversion and because one's budget constraint is affected by non-trivial charges in risk of death), people will not be willing to pay an amoupt 2X for a reduction in risk of 2 Δ P. Similarly, people's whose initial risk is Q instead of P, will generally be willing to pay something other than X for the same Δ P. We discuss some evidence about amounts people are willing to pay for different values of P and Δ P in Section IV.

11. J. Carlson, "Valuation of Life Saving," Ph.D. Dissertation, Harvard University (1963).

12. See, for instance, E. Crammond, "The Cost of the War," JOURNAL OF THE ROYAL STATISTICAL SOCIETY, Series A, Vol. 78, 361–399 (May 1915) or H. Boag, "Human Capital and the Cost of the War," JOURNAL OF THE ROYAL STATISTICAL SOCIETY, Series A, Vol. 79, 7–17, (January 1916). For a review of some relevant literature, see L. Dublin and A. Lotka, THE MONEY VALUE OF MAN, 1st and 2nd eds., The Ronald Press Co., New York (1931 and 1946) or D. Rice, "Estimating the Cost of Illness," AMERICAN JOURNAL OF PUBLIC HEALTH, Vol. 57, No. 3, 424–440 (1967). More recently, the livelihood-saving approach has been used in a number of governmental evaluation studies. See, for example, U.S. Department of Health, Education and Welfare, DISEASE CONTROL PROGRAMS: SELECTED DISEASE CONTROL PROGRAMS (1966a) and HUMAN INVESTMENT PROGRAMS: SELECTED HUMAN INVESTMENT PROGRAMS (1966b). B.F. Kiker ("The Historical Roots of Human Capital," JPE, Vol. 74, No. 5, 481–499 (1966)) and L. Thurow (INVESTMENT IN HUMAN CAPITAL, Belmont, California (1970)) have reviews of its general application to other areas of analysis. D. Rice and B. Cooper ("The Economic Value of Human Life," AMERICAN JOURNAL OF PUBLIC HEALTH, Vol. 57, No. 11, 1954–1966 (1967)) have one most extensively applied set of livelihood tables.

13. That is, if the earnings in year i are E_i, the probability of surviving until year is P_i, and the discount (or interest) rate is r, then the livelihood of a person n years old is

$$\sum_{i-n}^{\infty} \frac{P_i E_i}{(1+r)^{i-n}}$$

15. B. Conley ("The Value of Human Life in the Demand for Safety," AMERICAN ECONOMIC REVIEW, Vol. 66, No. 1 (45–55)) has recently argued that changes in expected present value of earnings provides a lower bound to individual willingness to pay for lifesaving programs. This conclusion requires a number of strong assumptions, however, on the nature of individual preferences and on a lack of interest by and for others in an individual's lifesaving valuation. Further, Conley recognizes that there is a range of income over which his conclusions do not apply. He assumes that this is at a very low level of income, but there is no evidence to support or to refute this assumption. P. Cook ("The Earnings Approach to Life Valuation: Reply to Conley," Draft Paper (1976)) suggests some illustrative values for the parameters of Conley's model which make it plausible that this will not be a lower bound for a large class of individuals.

16. Rice and Cooper, note 11 *supra*, and B. Cooper and W. Brody ("1972 Lifetime Earnings by Age, Sex, Race, and Educational Level," RESEARCH AND STATISTICS NOTE, DHEW (September 30, 1975)) have a widely used set of such tables.

17. The logical extension of the viewpoint which seems to motivate the livelihood precedure is to argue that an individual's consumption should be deducted from his earnings in calculating the value of his life—that his value is equal to the present value of the surplus he generates (note again the analogy with the slave). One implication of this "net livelihood" procedure is that society is made better off by the death of those whose expected net present value is negative—which is true of retired people and those who are near retirement, some of these receiving disability and public assistance payments, some children, and so on. Dissatisfaction with the implied judgment that society should not expend any effort to extend the lives of such people has led researchers to use income without excluding consumption: See, among others, R. Fein, THE ECONOMICS OF MENTAL ILLNESS, Basic Books, New York (1958); Klarman, note 1 *supra*, and M. Feldstein, COST/BENEFIT ANALYSIS AND HEALTH PLANNING IN DEVELOPING COUNTRIES, Discussion Paper, Harvard University (1970).

18. J.P. Acton, EVALUATING PUBLIC PROGRAMS TO SAVE LIVES: THE CASE OF HEART ATTACKS, The Rand Corporation, R—950–RC (1973).

19. Thirty-six of these respondents were selected at random from three communities in Boston (half men and half women); 19 were men in a trade union program, and 36 were in an advanced management program at the Harvard Business School. See Acton (note 18 *supra*, pp. 83–85) for a description of these samples.

20. B. Weisbrod, "The Valuation of Human Capital," JPE, Vol. 69, No. 5, 425–436 (1961).

21. H. Klarman, "Syphilis Control Programs," in Robert Dorfman, MEASURING THE BENEFITS OF GOVERNMENT INVESTMENTS, The Brookings Institution, Washington, D.C., 367–410 (1965).

22. Rice, note 11 *supra*.

23. M. Feldstein, note 17 *supra*.

24. For instance, we could examine the earnings of women with similar education and training who are employed full time in the market and impute those earnings to the women who stay home. See Posner, note 6 *supra*, pp. 79–80 for this opportunity cost argument.

25. J. Morgan, I. Sirageldin, and N. Baerwaldt, PRODUCTIVE AMERICANS: A SURVEY OF HOW INDIVIDUALS CONTRIBUTE TO ECONOMIC PROGRESS, University of Michigan, Survey Research Monograph 43, Ann Arbor (1966).

26. K.E. Walker, W. H. Gauger, "The Dollar Value of Household Work," Cornell University, New York State College of Human Ecology, Information Bulletin No. 60, Ithaca (June 1973).

27. Rice and Cooper (note 11 *supra*) assumed that all nonemployed women contributed a full share to home production and assigned the full-time earnings of a domestic worker to those women, about $2767 per year in 1964. They assigned no other value for household production to others. This implies, among other things, that it is frequently better to save women who do not work than it is to save women who work part-time. In Cooper and Brody (note 6 *supra*) the value of housework measured by Walker and Gauger (note 26 *supra*) was used, but no adjustment is made for men or for changed productivity after age 65.

28. Rice and Cooper, note 11 *supra*.

29. Walker and Gauger, note 26 *supra*.

30. Morgan et al., note 25 *supra*.

31. J.P. Acton, MEASURING THE SOCIAL IMPACT OF HEART AND CIRCULATORY DISEASE PROGRAMS: PRELIMINARY FRAMEWORK AND ESTIMATES, The Rand Corporation, R–1697–NHLI (1975).

32. Id., Sec. IV.

33. After this work was completed, Dorothy Rice (personal communication) informed me that the domestic worker's earnings for 1972 were about $4000. Resources did not permit recalculation of all the human capital tables to adjust for this fact, but we should note that it does not change the character of the methodological and empirical findings. If recalculated, the differential between men and women would increase during the working years and narrow somewhat over 65 years of

age. The average amount of willingness-to-pay measure would increase further over the human capital amount.

34. J. Dupuit, "On the Measurement of the Utility of Public Works," (1844) translation reprinted in READINGS IN WELFARE ECONOMICS, K. Arrow and T. Scitovsky, eds., R.D. Irwin, Homewood, Illinois (1969).

35. See, for example, P.A. Samuelson, "The Pure Theory of Public Expenditure," REVIEW OF ECONOMICS AND STATISTICS, Vol. 36, No. 4, 387–389 (1954) and "Diagrammatic Exposition of the Pure Theory of Public Expenditure," REVIEW OF ECONOMICS AND STATISTICS, Vol. 37, No. 4, 350–356 (1955); P. Bohm, "An Approach to the Problem of Estimating the Demand for Public Goods," SWEDISH JOURNAL OF ECONOMICS, Vol. 73, No. 11, 55–66 (1971); M.S. Feldstein, M.A. Piot, and T.K. Sundareson, RESOURCE ALLOCATION MODEL FOR PUBLIC HEALTH PLANNING: A CASE STUDY OF TUBERCULOSIS CONTROL, World Health Organization, Geneva (1973); L.B. Lave and W.E. Weber, "A Benefit-Cost Analysis of Auto Safety Features," APPLIED ECONOMICS, Vol. 2, No. 4, 265–275 (1970); E.J. Mishan, note 2 *supra*, and Zeckhauser, note 1 *supra*.

36. See J. Tobin, "On Limiting the Domain of Inequality," JOURNAL OF LAW AND ECONOMICS, Vol. 13, (October 1970); A.M. Okun, EQUALITY AND EFFICIENCY: THE BIG TRADEOFF, The Brookings Institution, Washington, D.C. (1975).

37. That is, effects that extend beyond the principal economic agent. A good example of externalities is the pollution that may be generated in the production of some goods. Neither the manufacturer nor the consumer of the good pay for the smoke (at least until recently), although a number of people experience the effects, would like to see them reduced, and would be willing to pay to have them reduced.

38. Dreze, in particular has argued the merits of using this procedure. See J. Dreze, "L'utilite Social d'une Vie Humaine," REVUE FRANCAISE DE RECHERCHE OPERATIONELLE, Vol. 23, 93ff (1962).

39. Thaler, note 1 *supra*.

40. R. Thaler and S. Rosen, "The Value of Saving a Life: Evidence from the Labor Market," paper presented at the NBER Conference on Income and Wealth, Washington, D.C. (November, 1973).

41. R.S. Smith, "Compensating Wage Differentials and Hazardous Work," study for U.S. Department of Labor (August 1973).

42. D. Usher, "An Imputation to the Measure of Economic Growth for Changes in Life Expectancy," in Milton Mose, ed., THE MEASUREMENT OF ECONOMIC AND SOCIAL PERFORMANCE, NBER, New York 193–225 (1973).

43. Acton, note 18 *supra*.

44. That is, risk of injury is probably positively correlated with risk of death. Omission of the first variable will bias the coefficient of the second variable away from zero, causing his estimates with the first data file to be too high.

45. Rice and Cooper, note 11 *supra*.

46. Advocates of this approach include T. Schelling, note 12 *supra;* V.D. Taylor, HOW MUCH IS GOOD HEALTH WORTH?, The Rand Corporation, P-3945 (1969); and J. Acton, note 18 *supra*.

47. Recently, a number of researchers have considered the nature of the utility function that may underlie an individual's willingness to pay for lifesaving. H. Raiffa (PREFERENCES FOR MULTIATTRIBUTED ALTERNATIVES. The Rand Corporation (1969) has shown under very general assumptions that a self-interested person, living alone (with no heir and a prepaid funeral), should pay more for a given reduction in probability of death if he is at a greater overall risk of death. J. Pliskin, M. Weinstein, and R. Shepard (UTILITY FUNCTIONS FOR LIFE YEARS AND HEALTH STATUS, Harvard School of Public Health, (October 1977)) and M. Weinstein, R. Shepard, and J. Pliskin (DECISION-THEORETIC APPROACHES TO VALUING A YEAR OF LIFE, Harvard School of Public Health (January 1975)) consider the valuing of life-years as a problem in multiattributed utility theory, where the joint or conditional nature of the "good" being offered makes a difference to the inferred value. P. Cook and D. Graham ("The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," Draft paper (1975)) explore the relationship between willingness to pay to avoid a loss and the compensation required to make a person as well off after a loss. M. Jones-Lee ("Valuation of Reduction in Probability of Death by Road Accident," JOURNAL OF TRANSPORTATION ECONOMICS AND POLICY, Vol. 3, No. 1, 37–47 (1969)) provides an analysis of the compensating variation required for various changes in the probability of death or injury. Usher (note 42 *supra*) and Conley (note 15 *supra*) formulate the issue as a life-cycle model in which the individual is assumed to try to maximize his expected lifetime utility, which depends directly on his consumption in each time period. Actual application is rare, however, as most writers have stopped with a theoretical treatment or have chosen an admittedly inferior technique for actual measurement.

48. J.L. Knetsch and R.K. Davis, "Comparisons of Methods for Recreation Evaluation," (1966) in R. Dorfman and N. Dorfman, ECONOMICS OF

THE ENVIRONMENT, W.W. Norton, New York (1972).

49. Acton, note 18 *supra*. Related work includes the survey of willingness to pay for selected disease entities conducted by M. Palmatier, "Willingness to Pay for Health Services: A Sampling of Consumer Preferences," Unpublished paper, Department of Economics, University of Southern California (January 18, 1969); a prototype survey for determining individual tradeoffs among attributes of disease reduction programs was developed by E. Keeler, MODELS OF DISEASE COSTS AND THEIR USE IN MEDICAL RESEARCH RESOURCE ALLOCATIONS, The Rand Corporation, P–4537 (1970). R.L. Berg ("Establishing the Values of Various Conditions of Life For a Health Status Index," in R.L. Berg, ed., HEALTH STATUS INDEXES, Hospital Research and Educational Trust, Chicago (1973)) and G.W. Torrance, D.L. Sackett, and W.H. Thomas ("Utility Maximization Model for Program Evaluation: A Demonstration Application," *ibid.*) have some imputed values for medical risk-taking based on the responses of physicians in their role as proxy decisionmaker for patients.

50. Part of the sample was a representative community sample in the Boston area, and part was a sample of young and middle-aged men in a business school program. A variety of questionnaire forms were used as it is not possible to report empirical results for the full sample of identical questions. The questionnaire for these surveys is contained in Acton (note 18 *supra*, Appendix).

51. Acton, note 18 *supra*, esp. pp. 92–105.

52. This finding is further evidence that individual preferences do not follow the implications of a livelihood-saving measure, which is strictly proportional to income. We can infer both risk aversion and an upper-limit of willingness to pay for a given mechanism of death reduction from these data.

53. That is, the responses to question types (2) were generally less than the responses to types (3), which were generally less than responses to types (4).

54. For instance, after thinking over what it might be like to be confined to a bed for a long period of time, his willingness to pay to avoid such disability might change.

55. E. Lindahl, "Some Controversial Questions in the Theory of Taxation," (1928), translated by E. Henderson; reprinted in R. Musgrave and A. Peacock, eds., CLASSICS IN THE THEORY OF PUBLIC FINANCE, 214–232 (1958).

56. Acton, note 18 *supra*.

57. Bolm, note 35 *supra*.

58. J.H. Dreze and D. de la V. Poussin, "A Tatonnement Process for Public Goods," REVIEW OF ECONOMIC STUDIES, No. 38, 133–150 (April 1971).

59. Bolm, note 35 *supra*.

60. P. Bolm, "Estimating Demand for Public Goods: An Experiment," Reproduced, Department of Economics, University of Stockholm (no date).

61. For instance, "you pay your actual maximum willingness to pay," or you pay some fraction, or you pay a proportion yet-to-be-determined, and so forth.

62. Other means besides a willingness-to-pay survey can be used to elicit the explicit values of individuals, but none of them answers the operational question of evaluation: How much should be spent on programs that change people's chances of death or disability? The exception to this assertion is a scaling technique that employs von Neumann-Morgenstern lotteries to determine a utility function. C.R. Neu demonstrates that this is formally equivalent to a willingness-to-pay approach ("The Use of Individual Preferences in the Public Valuation of Life and Health," unpublished Ph.D. Dissertation, Department of Economics, Harvard University (1975)). The remaining techniques cannot provide the operationally needed answer. For instance, a variety of psychometric scaling devices could be employed to measure people's attitudes toward attributes of program impact (say, death or disability), or their attitudes toward programs (say, heart attack ambulance or anti-hypertension programs). The results of such a scaling, however, do not answer the fundamental question of evaluation: Should scarce resources be committed? Suppose I know that Program A scores 8 and Program B scores 4 on a 10-point scale where 0 is very bad and 10 is very good. We do not know whether or not to undertake either program. Suppose we include information about program cost and define the status quo as 5 on the scale, we would still not know if either program should be undertaken. Furthermore, even if such a scaling produced an indication that a program should or should not be undertaken, the results are of limited applicability because we know only the valuation of a few programs rather than having a procedure that can be generalized. Another approach would be to ask people if they would like to see more, less, or the same amount spent on a given public program. If we then asked how much more should be spent, and specified the person's share of the cost, we would have a result equivalent to willingness-to-pay results and would answer the question of evaluation. Furthermore, if we ask enough questions, this iteration will produce a majority rule situation, which has significant appeal as a public decisionmaking criterion.

63. For instance, in Acton (note 18 *supra*) the conclusions as to net benefit of five interventions for out-of-hospital heart attacks were very similar under both methods of evaluation.

64. That is, if we were to tax away an amount up to the entire future earnings of individuals whose lives were saved, then we would cover the costs of such programs. In the absence of indentured servitude, we may not always realize even this situation.

65. Zeckhauser, note 1 *supra*.

Economic analysis and the Evaluation of Medical Programs

Jan Paul Acton, PhD*
The Rand Corporation
Santa Monica, California 90406

84

I. Introduction

Economics is the science of scarcity. It is useful in helping to answer such health program evaluation questions as:

- Should a new program be launched (e.g., should we add a mobile rescue unit with trained EMT's to an existing hospital emergency service)?
- Did we get our money's worth from a program that was started last year?
- Should we expand, contract, or eliminate an existing program?
- Should we expand our emergency medical program at the expense of another emergency medical program?
- Should we transfer resources from one non-emergency program to a particular emergency medical program (e.g., should the infrequently-used extra surgical suite be converted into an extra ambulatory care unit—or should it be the other way around)?
- Should we devote more of society's resources to emergency medical services and less to other social undertakings?

It is important to decide if the evaluation is *ex ante*—before a program is undertaken—or *ex post*—a retrospective analysis.

Economics is most helpful in analyzing the ex ante funding decision.

Cost-effectiveness Analysis: This is an efficiency criterion. It asks, is this the least costly way to achieve a particular effect?

Benefit-Cost Analysis: It asks, should the program be undertaken at all? That is, do the benefits outweigh the costs?

Benefit-cost analysis had four parts:

1. Predicting the consequences of a program—that is, assessing the probabilities.
2. Valuing the consequences or output—that is, measuring the benefits.
3. Assessing the costs of the program.
4. Selecting the best alternative.

II. Predicting the Consequences.

This is usually best done with the aid of a decision tree and the use of both objective and subjective probabilities.**

Major points to remember in assessing probabilities:

1. Most studies find that initial probability distributions are too narrow. Spread time out; admit it when you are uncertain!
2. Each person knows more about some topics than others. Don't spread the distribution too far when you do have a good basis for judgment.
3. Make use of different experts for different parts of the problem.
4. Most studies show that groups of people are much more accurate than a single assessor.

Suggested additional reading: Raiff's book is an extremely fine, and readable, introduction to how to be a practitioner of probability assessment.*

III. Valuing the Benefits

Three major alternatives exist:

1) Evidence from political process
2) Livelihood Saving (or Human Capital) measures
3) Willingness-To-Pay (or Individual Preference) measures.

Principal criticisms and comments about each include:

1. Political Process: Few consistent pieces of evidence on which to base evaluation. Implicit values range from a few hundred to over a million dollars per life saved.
2. Livelihood Saving: The most commonly-used technique in past studies. Widely

* The views expressed in this paper are the author's and do not necessarily represent those of RAND or any of its Corporate Sponsors.

**The decision tree is a display technique employed in decision analysis for decisionmaking under uncertainty. Howard Raiffa, *Decision Analysis: Introductory Declines on Choices Under Uncertainty*, Reading, Mass: Addison-Wesley, 1968 has a good introductory book. The handout material has an application in Jan Acton, *Evaluating Public Programs to Save Lives: The Case of Heart Attacks*, R-950-RC.

criticized because of the discriminatory treatment of women, retired persons, those who do not work, and those who will not reach working age.

3. Willingness-to-pay measures are based on the premise that individual preferences should count in programs that affect people's lives and happiness. Some work has been done based on implicit valuations—for instance in extra hazard pay—but considerable variability is observed. Preliminary evidence suggests that people can respond well to survey-type questions and yield useful information, but additional work is needed.

These alternatives are discussed and critized in detail in the attachment by Jan Acton. Measuring the Monetary Value of Lifesaving Programs, P-5675.

IV. Selecting the Best Alternative

Major points to remember:

Don't use a benefit-cost *ratio* to choose. Select the alternative with the greatest *net benefit*.

Don't just select the alternative with the greatest reduction in mortality rates. Remember, changes in mortality rates may be more important for some groups than for other groups of people.

Check for sensitivity to assumptions and data used?

- Would the choice change if slightly different measures of benefit were used?
- Would the choice change if the probabilities were somewhat different?
- Would the choice change if the alternatives available are slightly different?

If yes to any question, then try to sharpen the data or values used.

Check for omitted factors and variables that might tip the balance the other way. If the decision seems sensitive to these omitted elements, try to incorporate them formally in the analysis.

ILLUSTRATION of Decision Analysis applied to the evaluation of two new programs for an existing emergency service. These assumptions are admittedly arbitrary and somewhat unrealistic, but they illustrate the methodology.

Assumptions:

- Two programs are available: one for treating heart attack victims, one for trauma victims. Only one can be selected. They cannot be combined.
- The outputs of both programs consists mainly in reducing the number of people dying. Other outputs are not important. The program will apply to a population of 10,000 people.
- Both programs reduce the probability of death by 50% for those eligible people reached.
- The probabilities of death, of calling for the program, and of being treated successfully are independent for each program.
- Heart attack and trauma events occur independently.
- The probability of calling the heart attack program, given a heart attack, is 50%.
- The probability of calling the trauma program, given a trauma event, is 80%.
- The heart attack program will be able to reach and help 80% of those who call.
- The heart attack program costs $100,000 per year.
- The trauma program costs $70,000 per year.
- The trauma population is younger and has a better prognosis if "saved" by the program. In the range of expected effectiveness expected, each person is willing to pay an average of $8 per year for each chance in 10,000 that the program reduces his chance of death.
- The heart attack population is somewhat older and has a worse prognosis if "saved" by the program. In the expected range of effectiveness, each person is willing to pay an average of $3.75 per year for each chance in 10,000 that the program reduces his probability of death.

Conventions:

We will designate points in the decision process where a choice must be made with a square.

Chance nodes are indicated by a circle.

Costs associated with action taken are indicated by a barrier across the pathway.

Figure 1: Current Situation, No New Program

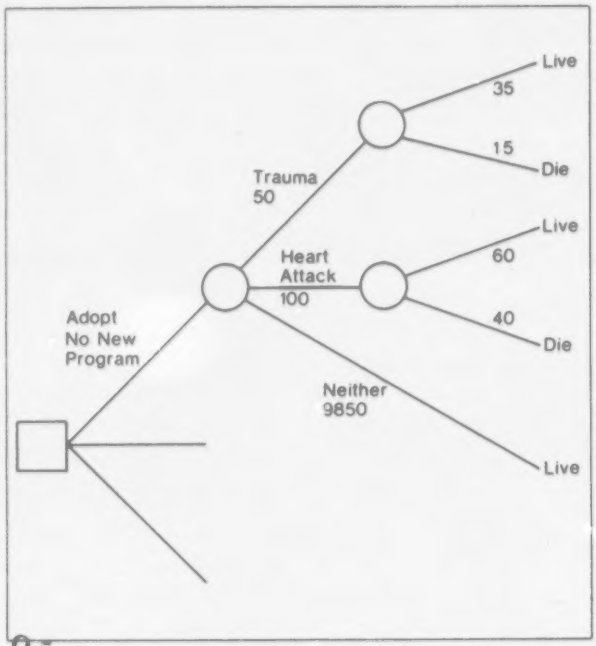


Figure 2. Effect of Heart Attack Program

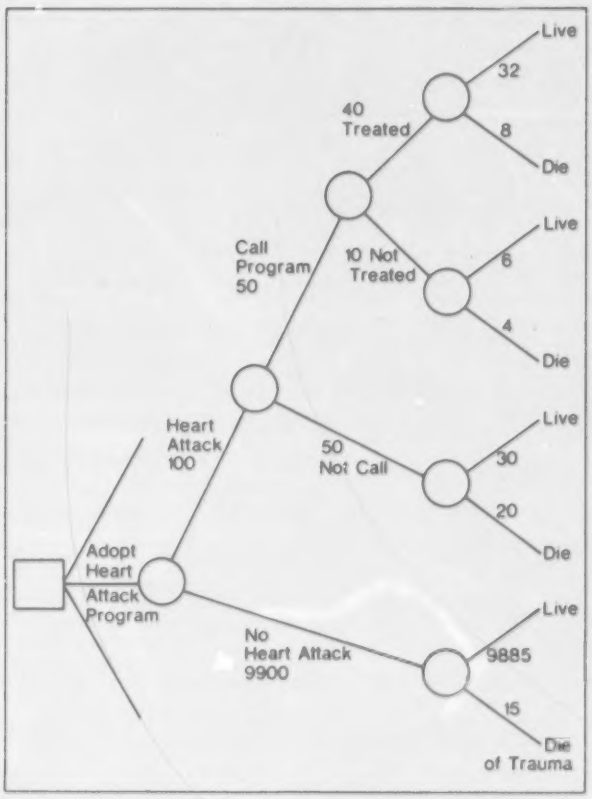


Figure 3. Effect of Trauma Program

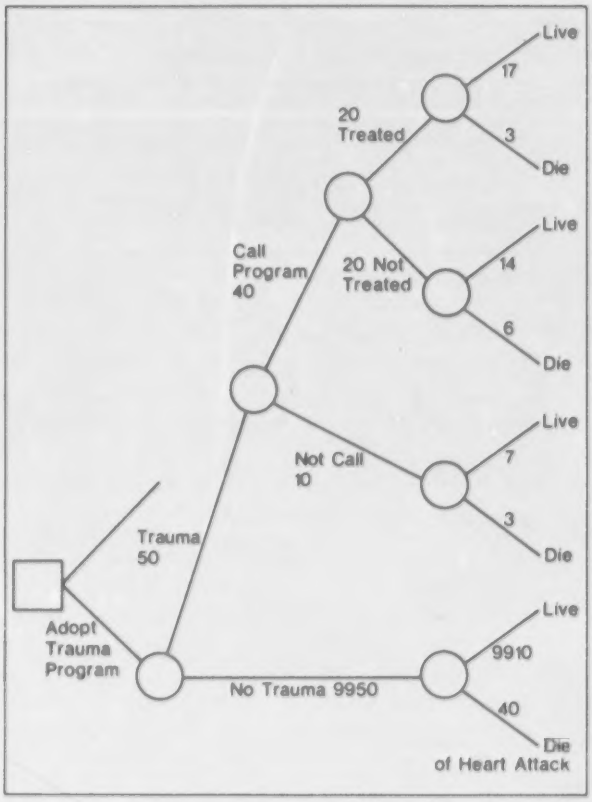
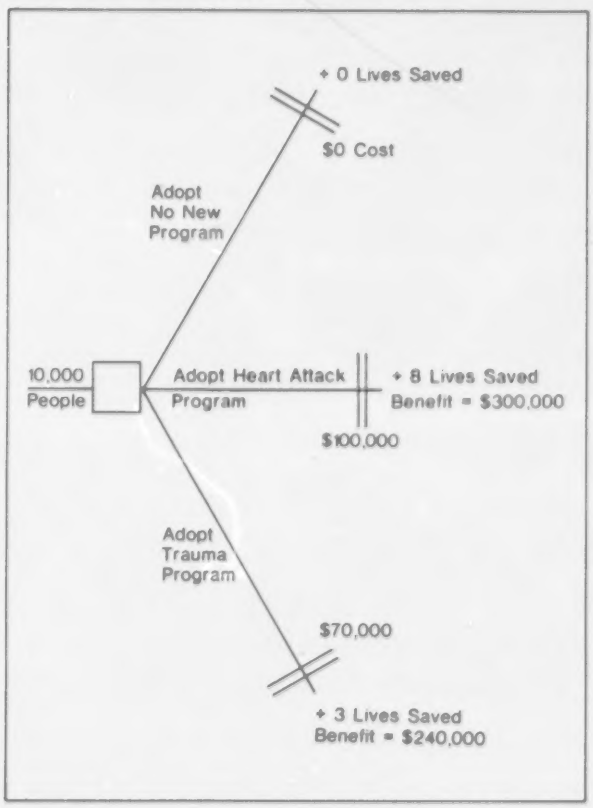


Figure 4. Decision Alternatives.



Appropriateness and Feasibility of Randomized Field Tests

Robert F. Boruch
Northwestern University

When it is proposed to persons working in various service delivery systems that their programs should be evaluated by experimental methods, strong doubts are almost invariably expressed about the feasibility or even possibility of experiments in public service delivery systems. Robert Boruch, an evaluation methodologist, has identified hundreds of experiments carried out in just such settings. This paper summarizes his experiences and views. It presents a strong rationale for evaluation, an overview of problems and methods involved in program evaluation, and the case for conducting randomized experiments.

1. Introduction

This paper reviews briefly what we have learned about appropriateness of mounting field experiments to plan and evaluate social programs and about the feasibility of such tests. "Appropriateness" is considered here as a kind of precondition for feasibility, one which exercises a direct impact on the level and nature of a subsequent feasibility study. Feasibility here concerns those conditions which enhance or detract from the successful conduct of an experiment. This discussion depends heavily on studies of efforts to foster the use of randomized tests of programs in field settings. We adhere to the following outline:

2. Appropriateness of Evaluation and, in particular, of Randomized Field Tests
3. Historical Precedent as a General Test of Feasibility
4. Pilot Feasibility Experiments as a Test of Feasibility
5. Direct Constraints on Feasibility of Randomized Tests

The Bibliography attached provides some background support, in the form of field tests actually mounted, for the opinions offered here.

2. Appropriateness of Evaluation and, in Particular, of Randomized Experiments

Several questions generally need to be answered before an experiment is considered much less mounted. The answers to them serve not only as guides in deciding whether and what to evaluate, but also determine subsequent feasibility of an experiment. Those questions, discussed very briefly in the following remarks, include:

2.1 Is there any interest in evaluation, much less an experimental test?

2.2 Is an impact evaluation rather than some other type appropriate in the setting at hand?

2.3 Are the effects of the program currently debatable?

2.4 If so, what is the proper standard for an impact evaluation?

2.5 Will methods other than a randomized experiment suffice for impact estimation?

2.1 Interest in Evaluation

If sponsors of a program have an interest in obtaining a fair appraisal of a program's effectiveness, relative to any standard, then mounting an evaluation, randomized or otherwise, is considerably more feasible. It is doubtful, for example, that Career Education programs supported by the National Institute of Education would have or could have evaluated themselves without encouragement *and* demands made by the agency. That sponsor's support is an insufficient interest source is also clear from cases in which despite sponsor demands, rigorous evaluations have been subverted by program staff.

So program staff and developer interest is also a determinant of feasibility of any evaluation. Reputable program developers will often agree that an evaluation is necessary as will program staff. But assuring that the interest is not honorific is altogether a different matter. Some strategies for assuring cooperation with staff must be worked out beforehand. Some of these are discussed in Riecken et al. (1974) and in Section 5, below.

The client population usually has some vested interest in the outcome of an evaluation. And this interest is most often exemplified in ways other than active collaboration in the rigorous process of evaluation. Often, the difficulties in experiments turn around the randomization process. Tactics for determining and enhancing feasibility are discussed in Section 5 for the particular case of experimental evaluations.

If there is no interest in a fair estimate of program quality from any of these quarters, then an evaluation, experimental or otherwise, is likely to be of little use to anyone except the individual conducting the evaluation. If there is active opposition from one or more of these quarters, matters become difficult indeed.

2.2 Type of Evaluation

The "evaluation of a program" often implies a disparate array of activities. And to avoid needless confusion, one ought to recognize the legitimacy of several functional categories of evaluation:

evaluation of program *objectives*
evaluation of program *process* or *operations*
evaluation of *cost/benefit* ratios
evaluation of *impact*.

Each of these is related to more elaborate taxonomies of evaluation activity generated by Federal agencies (e.g. U.S. AID) and especially by academic researchers (e.g. Stake). The taxonomies are a useful guide through the thicket of types and methods, but we focus on only four here for simplicity's sake.

The first category, evaluation of program objectives, involves pinning political, ethical, or social values to the announced goals of the program. Typically, this type of evaluation is tied to real or imagined needs of a target group; it is implicit in most policy development and policy criticism; and it is based on information which purports to show that there is a social problem and that a program is one way to ameliorate the difficulty.

The second category, evaluation of program activities, involves determining whether and how well some well specified standards for implementing the program are met. This class of activities is often managerial in its orientation, addressing questions such as: Is there a clearly specified product being developed? Is the product or service being offered to the proper target group? Is the product accepted and, if so, to what extent? How much does the program system cost? A second major perspective is also relevant here and is more technological in character. The expert program practitioner or judge may ask whether the program's elements and conduct are consistent with the state of the art in the relevant discipline, and whether there are any remarkable inconsistencies, nonuniformity, or deviations. The standards here are those of a discipline, firmer perhaps than the attachment of social or political values to program goals; they focus on the immediate scientific common sense of a program rather than on ultimate outcome, and that too is important.

The third class of common activities is the cost/benefit analysis. This covers a variety of sins, but most often involves assuming that there is indeed a program benefit and assuming that the benefit has some value. The costs have been traditionally a bit easier to pin down. The objective, given those assumptions, is to provide criterion for determining how scarce resources ought to be allocated when there are numerous competing demands for those resources. Again, whether the program does indeed have an effect is generally assumed, taken for granted, or judged relative to a traditional standard for which there is some consensus.

The final form of evaluation and the one which interests us most here concerns the relative effects of the program, i.e. impact evaluation. It attempts to answer questions such as the following: Which of two education programs enhances student achievement or ability or attitude most? Does a new surgical treatment have fewer side effects than the current one? Which of several health education programs has the largest effect on actual health status of individuals or cities or regions? In each case, one asks how the program, or service, or delivery mode, works with respect to some standard or alternative.

Each category of evaluation is legitimate and important. And, of course, nothing prevents each type from being conducted simultaneously. Indeed, most major program evaluations include features of each type. The first and third categories are generally more feasible than the second and the last. But the information they provide differs in each case. Whether one or another category is most appropriate depends heavily on the interest of the principal sponsor of an evaluation.

2.3 Evidence on Impact

If a program's effect on a target population is already known to be positive and its magnitude and cost are similarly well established, there appears to be very little point to conducting an impact evaluation, whether randomized or not. Studies undertaken for strictly scientific reasons, rather than for the sake of policy planning and development, are an exception and this case we put aside for the moment.

In most instances the need for an evaluation arises because there is some honest disagreement among experts about the nature of an effect. The lack of agreement or even of informed opinion may stem from the fact that the program is a completely novel one, as many innovative social programs are advertised to be. Or the disagreement may stem from previous research which permits only the most equivocal of inferences about the nature of a treatment's effect. The Negative Income Tax Experiment, for example, was mounted because regression, covariance, and other correlational research techniques were insufficient for supporting major policy decisions; the effects of various levels on income subsidy on work behavior and so on could not be predicted with sufficient accuracy or with a sufficiently low level of am-

biguity. Similarly, equivocal data accumulated over the past 15 years has led to the development of the current national clinical trials to test experimentally the effect of special diets and drugs on arteriosclerosis.

Disagreement here applies not only to the program itself but also to the manner of its delivery. It is well known, for example, that certain nutritional supplements have positive and detectable effects on physical development of children. But how to manufacture, deliver, and encourage acceptance of such supplements among malnourished children in depressed regions is often not at all clear. The agreement of judges that evidence on best methods of delivery is scanty serves as a justification for impact evaluation, including randomized tests and alternative methods of delivery and encouragement.

Similarly, disagreement may occur about components of a program rather than about the total program. Alternative methods of screening individuals, of training service delivery staff, of referral service staff, or of program recipients, and so may not be central to a complex program, but may indeed warrant impact evaluation.

2.4 The Standard and Impact Evaluation

Two kinds of standards are pertinent in deciding what type of impact evaluation is appropriate, and in settling on a randomized experiment as the design of choice. The first kind concerns standard against which estimates of impact should be judged. The second concerns standards for judging the equivocality or bias in estimates of program effects.

Standard for magnitude estimation. One can of course choose a historical precedent to gauge the impact of treatment. In the ideal case, one has a long stable time series available, the program is introduced abruptly, and the program effect is gauged by its effect on the time series. There may be other similarly ideal empirical ways to specify null conditions—how things are in the absence of any extraordinary program effect. They may include naturally occurring, entirely equivalent comparison or control groups.

Or, the standard against which effects are judged, the null condition, may also be specified by assumption or by fiat. In the former case, for example, one might be willing to assume, based on theory, commonsense, or whatever, that there will be absolutely no improvement in the condition of a mentally retarded group without a program. In the latter case, one might specify, as Nixon did, that if a crime reduction of 10% occurs, then the program (whether it is really in the field or not) will be declared a success.

Now any of these standards may, in particular instances, be quite appropriate. Enough may be known from theory to specify the null condition quite accurately. There may be sufficient theory and data to specify the baseline standard well. And in some instances, the use of these options is fine.

The problem, however, is that in social program evaluation, neither theory nor prior data are sufficient for specifying null conditions adequately, for assuring that the supposed standard of comparison is a fair one. Furthermore, even the theory which does exist may be insufficient for coping with the competing explanations for the finding that an effect is significant. The effect found may stem from influences completely outside the program, it may have been a continuation of an unrecognized trend, and so on.

The randomized experiment is, in this context, most appropriate when null conditions cannot be prespecified well from prior data, by assumption, or by fiat. That is, it sets up a timely comparison group whose equivalence to a treated group is guaranteed in the long run and which can be used as a fair and reasonable benchmark for estimating program effects. The experiment also reduces the equivocality problem notably: The number and plausibility of competing explanations can be reduced.

The standard for equivocality of inference. The benefit of randomized experiments is that if they are conducted properly, the judgements one can make about existence and size of effect are less susceptible to attack. That is, other methods may produce an estimate of program impact which is susceptible to bias, due to unrecognized influences, extraneous factors, and so on. There is a fine state of the art in identifying competing explanations for findings derived from observational (nonrandomized) evaluations, and it will not be discussed here. See, for example, Campbell and Stanley's (1966) classic monograph or a revised edition, Cook and Campbell (1976).

There exists, however, no formal technique for attaching a "level of equivocality" to the findings from quasi-experimental studies. Whether such a system could be drawn up depends heavily on the particular substantive area and on whether the competing explanations are plausible or realistic. Establishing the tenability of the last time, regarding realism, brings us to another criterion in establishing the appropriateness (and consequently, feasibility) of randomized experiments: Can methods which do not rely on randomized assignment yield estimates of program effect which are close to those which one might obtain in an experiment? Some tentative answers to the question are given in the next section.

2.5 Possibly Suitable Alternatives to Randomized Trials

The basic idea here is that one ought to determine if *randomized experimental tests are unnecessary because we might be able to use a variety of quasi-*

experimental and (or) algebraic adjustments to obtain unbiased estimates of program effect. The exact conditions under which a randomized experiment will yield the *same* estimate of program effect as a nonexperiment are, in principle, specifiable beforehand. However, determining whether those conditions are actually met in the field is usually difficult and often impossible. One simply does not know whether the analytical conditions assumed for the nonrandomized evaluation and analysis hold in reality. Consequently, many such evaluations cannot be used to support contentions about program impact. That the problem is a persistant one is evident from reviews by Wargo and his colleagues (1971) and by Bernstein and Freeman (1975) of evaluations of Federally-subsidized social programs: In the majority of the nonrandomized evaluations, there were competing explanations for the findings, explanations which could not be ruled out on common sense grounds or on the basis of the empirical data collected in the evaluations.

To get an empirical fix on the matter we can try an approach geared chiefly toward understanding the limits of statistical manipulation. Here, one locates (or conducts) a randomized experimental test of a program, and in addition, collects sufficient nonrandomized data to support ostensibly appropriate quasi-experimental assessment of the same program. Suppose, for example, that data are obtained on individuals who have been randomly assigned either to a treatment program (T) or to a control condition (C). Similar data are also collected on an additional group (C') whose members, though not randomly assigned, are regarded as members of the C group and to the T group prior to treatment. The question is then posed: How does the estimate of program effect based on ordinary analysis of variance of the T-C group compare with an estimate of effect based on the T-C' groups and conventional statistical techniques such as matching covariance analysis, or change scores analyses? The answer is important insofar as it helps us to understand the nature and direction of bias that may be obtained when using techniques such as covariance analysis purportedly yield unbiased estimates of effect without randomization.

That estimates of effect will often (but not always) be biased if we rely solely on nonexperimental evidence becomes obvious with some concrete examples. Consider the simplest form of nonexperimental analysis—comparing the condition of program recipients before the program's introduction to their condition afterward. This before-after (or pretest-post-test) approach is common despite the fact that any increase or decrease in average condition may be entirely attributable to unrecognized growth or development processes.

In the Michigan arthritis studies, for example, severity of condition *increased* after the introduction of an arthritis treatment program. Based on this information alone, we might erroneously conclude that the program's effect was negative, i.e., it actually harmed program participants. In fact, we know from randomized experimental tests that the equivalent control group's condition deteriorated even further, and consequently, the proper inference is that the program did indeed have a beneficial effect. (See Deniston & Rosenstock, 1972).

In before-after evaluations of compensatory education programs, cognitive scores may increase, decrease, or remain stable. The change tells us virtually nothing about the program impact simply because we usually do not know for the subgroup tested and for the particular test what the change would have been in the absence of the program. (See Wargo et al., 1971.)

Usually one attempts to find a comparison group against which to gauge the condition of program participants, and also to reduce the equivocality underlying most before-after designs. But this is also hazardous to the extent that the comparison group differs systematically and often in unknowable ways from the participant group.

For example, one facet of the Salk vaccine trials involved comparing volunteer vaccine recipients to an allegedly equivalent, "natural" comparison group of nonrecipients. The vaccine's effect in this nonrandomized quasi-experiment was positive. But estimates based on a second facet of the trials—randomized tests—gave estimates of effect which were 14% higher than the value based on the nonrandom tests. Given only the evidence from the nonrandom groups then, we would have concluded that the vaccine was notably *less* effective than it actually was in reducing polio incidence (Meier, 1972).

In randomized tests of a retardation rehabilitation program, Heber et al. (1972) collected data on an additional plausibly equivalent comparison group—siblings of children enrolled in the program. The difference in observed IQ between program participants and nonparticipants in the randomized test was about 36 points. A comparison of program recipients against their siblings (an ostensibly equivalent contrast group) yielded a 45-point difference. Had we relied solely on the "natural" comparison group, we would have overestimated the program's impact in this instance.

At this point, the statistically knowledgeable and critical reader might observe that there are algebraic techniques which purportedly "adjust out" differences between groups and which equate

groups which differ initially, in order to avoid biases such as these. The techniques—matching program participants and nonparticipants with respect to their demographic or other characteristics, covariance or aggression analysis—are sophisticated but do require strong assumptions about the underlying nature of the data. More importantly, those assumptions may not be an adequate picture of reality, i.e., of how individuals will behave in the absence of any program intervention. To be specific, when groups differ initially and the difference persists, then these methods will *not* perform adequately if the matching variables or covariates are measured imperfectly or incompletely. Some of the more advanced techniques accommodate the problem of fallible measures reasonably well, provided that reliability of the data is not too low (e.g., Porter, 1967). But none accommodates the specification problem satisfactorily: in many cases, we are very likely to leave out variables which are important but which are unmeasured or unmeasurable. In either case, the adjustment process is imperfect, and estimates of program effect will often be biased. How often will they be biased? It is impossible to say, but a few examples may help to illustrate the problem.

> In the Michigan Arthritis Study, a comparison group was identified, differences between this group and program participants were reduced by matching individuals, and estimates of program effect obtained. The estimates of effect based on this comparison is near zero; that is, despite selection of a matched group, the estimate obtained by comparing these individuals to the program participants is biased, relative to the estimate obtained from the completely randomized data (Deniston & Rosenstock, 1972).
>
> The Middlestart program was designed by Yinger, Ikeda, and Laycock (1967) as a special pre-college program for promising high-school students. In their original evaluation, some students were assigned randomly to participant and control groups. Others were assigned on the basis of post-facto matching. That is, five sets of treatment and comparison groups were constructed; they were not randomized and were equivalent only in the sense that they were matched on the basis of their demographic characteristics. If one examines the pooled data, one finds a significant difference of about six months in grade equivalent achievement test scores between participants and nonparticipants. However, if one examines only the randomized set of students, the estimate is far lower and quite negligible. In this case, the nonrandomized comparisons yield estimates of effect ranging from zero to a two-year difference in achievement test scores (Boruch, Magidson, Davis, 1976).

Time-series designs are also a promising approach to estimating program impact. Here one observes some outcome variable over time (e.g., rape rate over the last three years), introduces the program, and then tries to detect subsequent change in the variable (e.g., a drop in incidence of rape). The time-series approach is promising to the extent that there is no good competing explanation for the change in the outcome variable, such as changes in the accuracy of measuring the incidence of rape, and to the extent that the time series is suitable, so that a discontinuity will be obvious if it occurs. That time-series analysis is often not possible and that it will often yield estimates which differ from those based on experimental evidence is also clear, however.

> Considering the Cali (Colombia) evaluation of nutrition and education programs, we find that an estimate of program effect based on short time-series projection from the control group is biased downward drastically. The time-series estimate of effect on children's cognitive skills is half the size of the effect based on test scores of randomized recipient and nonrecipient groups. The bias would be smaller if a much longer time-series had been available (see McKay, McKay, & Sinnesterra, 1973).
>
> Time-series data on polio incidence prior to the Salk trials were insufficiently valid and comprehensive to support credible time-series estimates of the vaccine's impact. Similarly, novel programs such as the Career Education Projects supported by the National Institute of Education, the Headstart variations efforts of the U.S. Office of Education, and others could not be evaluated on the basis of time-series analysis simply because valid, stable time-series data on important outcome variables is unavailable.
>
> In the Michigan Arthritis Study, time-series estimates of effect were 10% higher than estimates based on randomized experimental tests in the same populations.

Of course, there have been studies employing much less competent methodology than even the imperfect ones we have described which have also led to erroneous conclusions. The more dramatic examples have occurred in medicine, where medical or surgical remedies, adopted on the basis of very weak evidence, have been found to be of no use at best and to be damaging to the patient at worst.

> The so-called frozen stomach approach to surgical treatment of duodenal ulcers, for example, was used by a variety of physicians who simply imitated the technique of an expert surgeon. Later experimental tests showed prognoses were good simply because the

originating surgeon was good at surgery and not because his innovation was effective. It provided no benefit over conventional surgery (Ruffen et al., 1969).

Anticoagulant drug treatment of stroke victims had prior to 1970 received considerable endorsement by physicians who relied solely on informal observational data for their opinions. Subsequent randomized experimental tests showed not only that a class of such drugs had no detectable positive effects but that they could be damaging to the patients' health [see Hill et al. (1960) and other examples described in Rutstein's (1969) excellent article].

92

None of this should be taken to mean that estimates of program impact based on experiments will always differ in magnitude from those based on nonrandomized assessments. The estimators will be close, for example, if there is no systematic difference between characteristics of the individuals assigned to one program variation and those assigned to another. If in a particular research project there is no systematic association—i.e., there is a kind of natural randomization process—or if such differences can be removed statistically, then we may expect various types of designs to produce similar results.

We have been able to document few instances of this occurrence, however. The first stage of Daniels's evaluation of the DANN Mental Health program, for instance, involved allocation of incoming patients to the experimental treatment ward on the basis of number of beds available in each. Controlled (deliberate) randomization was introduced after ward turnover rate had stabilized. Comparisons of the characteristics of ward entrants prior to their treatment in the first nonrandomized stage to the characteristics of entrants admitted in the second (deliberately randomized) stage showed no important measurable differences between the groups. More importantly, separate analyses of the nonrandomized and randomized groups yielded very similar estimates of program effect.

An essential condition for similarity of estimates is that prior to program introduction, there be no systematic association between characteristics of eligible program candidates and their participation in the program. The association may be slight enough at times to give us some confidence that the program effect is in the proper direction even if we recognize that the magnitude of the estimator is likely to be in error. Holt's (1974) evaluative studies of sentence reduction in prisons is informative in this respect. A number of nonrandomized studies on early versus late release of individuals from prison suggested that length of sentence (within certain limits) had no impact on post-prison behavior. Later randomized experimental tests demonstrated that the direction but not the magnitude of the early estimates of the effect of early release were appropriate.

In each of these cases, as in others (see Boruch, 1975), randomized tests were needed to verify that unobserved influences were not entirely responsible for the results obtained in nonexperimental studies. More specifically, the Daniels experiment helped to rule out the possibility that program effects estimated from the nonexperimental data of the first stage were attributable to subtle differences in patients assigned to each ward rather than to the ward program itself. In the Holt work, the experiment helped to demonstrate that the success of early releases was not entirely attributable simply to very expert judgments by parole boards about the likelihood of a parolee's returning to prison, but that the length of sentence actually has no discernible effect on recidivism within certain limits.

Remarks. It is clear that in some nonrandomized evaluations attempts to statistically "adjust out" preexisting differences between treatment and nonequivalent comparison groups can lead to biased estimates of the treatment effect. The direction of these statistical biases in certain stereotypical cases can be such that the treatment will appear to have had a *negative* effect. Biases of this sort probably underlie some evaluators' declarations that Headstart programs and Manpower Development and Training Act Programs had a detrimental effect on program participants. Some of the conditions under which the statistical biases may appear are described, along with examples, by Campbell and Boruch (1975) and Boruch (1975). To better gauge the extent to which new statistical approaches to analyzing nonrandomized data actually avoid this problem, the Project on Secondary Analysis (Boruch, Wortman, & DeGracie, 1975) is applying competing methods of analysis to the same data set and documenting the biases underlying each method.

Other researchers are conducting investigations along related but distinctive lines. That research is often supported by special divisions within Federal agencies—NIE's Program on Measurement and Methodology (Porter, 1975), HEW's National Center for Health Services Research AID's Division of Methodology (Technical Assistance Bureau)—which are designed to foster methodological investigations and which should help to identify approaches to evaluation which have far fewer technical weaknesses and greater flexibility in the field than those currently available.

3. Historical Precedent as a General Test of Feasibility

The idea that experiments are an ideal but impractical method for estimating relative program effects is often proposed. But the contention

about impracticality is rarely supported with any evidence or analysis. In fact, just a little homework can yield a good deal of information about experimental tests which have been mounted. And that information can be used at least as contextual or background evidence for making a crude judgement about feasibility of experimental tests on the program at hand.

The Bibliography of this paper, for example, contains excerpts from a list of more than 200 experiments (Boruch, 1974) and illustrates the remarkable variety of social programs which have been subjected to experimental field test. In the economic arena (Section IX), for instance, the Negative Income Tax experiments, the Housing Allowance Experiment, and the Health Insurance Experiments represent remarkable efforts to determine the best of alternative economic subsidy plans. There have been dramatic judicial experiments (Section II) which demonstrate the feasibility of randomized appraisals of the effectiveness of changes in judicial rules and practices. Experiments have been successfully mounted to assess the effects of police training programs (Section II), rehabilitation programs for juvenile and adult offenders (Section I, II, XII), and programmatic decisions have been based on the results of these. Socio-medical (Section XI) and mental rehabilitation experiments (Section III) are represented here and abroad. Educational experiments are quite common, and although most are rather small, the Cali (Colombia) experiments on compensatory education for nutritionally deprived children, the research on "Sesame Street" and "The Electric Company" in television-based education, and at least a dozen others involve sizable samples, complex programs, and high-quality investigation (Section IV). There have been a large number of experiments conducted to identify superior methods of assuring quality and completeness of information transmission in audits and surveys (Section VI); most have been designed in the broader context of Federal data-collection efforts, and they provide good evidence for choosing ways to accomplish part of that mission. Because some experiments which take place in industrial settings are relevant to groups often targeted for social programs (the aged, the poor), illustrative experiments in this context have also been included (Section VIII).

Experiments vary in other ways. Some experiments, for example, have been conducted to estimate the impact of important, very *small* elements of a very complex treatment; e.g., laboratory research on the most effective size of letters and numbers in television broadcasts was conducted prior to large-scale evaluation of the more complex total Electric Company program. Others, like the Negative Income Tax Experiment, involve more simple and homogenous "programs"—the provision of income subsidy, the administration of a rule, etc. There is a surprising variety in the target of randomized assignment: children, in assessments of many education programs; adults, in all substantive program categories; families, in economic experiments; neighborhoods, in fertility control and communications experiments; hospitals, school districts, and others. Many of the references cited in the Bibliography reported in only one experimental test in a series of simultaneous replications, as in the Negative Income Tax Experiments (in Wisconsin, New Jersey, Indiana, and Colorado). A series may consist of a sequence of experiments and quasi-experiments, dedicated to long-range development, testing, and revision of a program. The Goodwin-Sanders (1972) work exemplifies this last strategy; it involved sequential assessments of tape-playing devices for education, used on school buses enroute to the children's homes.

Some experiments have not been implemented completely, of course. For example, the Hornik et al. (1973) assessment of television education in El Salvador was designed in part as a randomized experiment, but the randomization procedure failed in the face of what appear to have been insurmountable administrative difficulties in the evaluation. Similarly, efforts to conduct randomized tests have at times been unsuccessful in assessments of delinquency programs (Clarke & Cornish, 1969), education programs (Owens et al., 1974), and elsewhere. Still, many experiments have been mounted successfully by designing the study to accommodate political and social factors which might otherwise undermine randomization and valid measurement: for example, the Manhattan Bail Bond experiments, which conflicted with the vested interests of bail bondsmen; experimental tests of nurse-practitioner programs by Sacket (1973) which conflicted with the interests of some physicians, and others. In fact, most of the experiments listed in Bibliography did accomplish planned randomization.

Outright failures of randomization undoubtedly occur more frequently than the Bibliography suggests, and, of course, the reasons for failure are important. The only systematic analysis of those reasons available so far, however is Conner's (1974) set of case studies and our own analysis. Conner identified the directness of the evaluator's role in the randomization process as a key ingredient of success. Other ingredients are important, but the current scarcity of documentation on failures, aside from evidence provided here, makes identification of reasons for failure difficult. Hornick and others have displayed an exceptional willingness to examine the reasons for unsuccessful randomization, and to build on that information to develop better methods of analyzing the resultant observational data.

Given the number, quality, and variety of field experiments which we have been able to identify, the *general* contention that experiments are impractical is a bit underwhelming. There are, however, some other important feasibility issues which have also been used to justify not randomizing. The more typical ones are outlined in the other sections of this paper.

Remarks. That a notable number of randomized experiments have been mounted in the field does not demonstrate the feasibility of experimental tests under all or even most social conditions, of course. The examples do, however, serve as valid evidence against the broad contention that rigorous appraisals of the effect of a social program are rare or impossible. They also serve as a basis for examining conditions under which controlled tests appear to be most readily mounted. For example, many such tests compare the effects of various material products, such as two different income subsidy plans, different drugs, different sets of written instructions, and so on, rather than the effects of social programs which are based heavily on personal skills of program staff, such as two rehabilitation programs for the mentally ill. It is conceivable that experimental tests of the latter sort are more difficult to conduct because we do not know enough about designing tests which are especially sensitive to staff skills or which do not threaten the status of program staff. Similarly, many experiments involve estimating the effects of *new* social programs, while relatively few are devoted to ongoing programs. That strong traditions, beliefs, and ingrained practice common to ongoing programs are less conducive to planned evaluations, has been recognized by legal researchers (e.g., Hans Zeisel), by medical researchers (e.g., Thomas Chalmers), and others. But this is not to say that experimental tests of less material programs, or of ongoing programs are undesirable or impossible. It is to say that considerably more effort must be expended in mounting experimental tests of ongoing programs and that the efforts may not pay off in a successful test if regular program staff resist the idea of evaluation.

A different reason for failure of an experiment concerns the public's rejection of an unfamiliar idea—randomization. Some good experimental tests have been undermined by premature and naive acceptance of randomization as well as by premature and naive rejection. Public education is likely to help make acceptance more informed. But in addition, some empirical work by program evaluators on related determinants of acceptability can be justified. Hendricks and Wortman (1974), for example, are examining the effects of a program candidate's assigning himself randomly to program condition, because assignment by program staff or by an impersonal institution appears at times to generate resistance. These small laboratory experiments and case studies such as Conner's (1974) are likely to be helpful in generating more realistic approaches to handling the problem in the actual field experiment.

We consider the matter of randomization in more detail in a later section of this paper.

4. Pilot Feasibility Experiments as a Test of Feasibility

The suggestion just made, that examining precedent can be helpful in making crude judgments about the feasibility of an experiment, is a reasonable one. But it is considerably less direct an approach than is generally necessary. Now one relatively *un*common but quite direct approach to the matter is to mount a live pilot experiment, a little field test to appraise feasibility of the full-blown field experiment.

Such a pilot feasibility study may be a unified endeavor as we've implied, a dress rehearsal before a live but very limited audience prior to the main test. This is not a common tactic in the social sciences where the exuberance of a youthful science and short time frame may prevent its more frequent use. But it is not uncommon in other arenas, including medical experimentation. The more common approach, of course, is to set up a number of small tests or studies prior to the main study to assure feasibility of special features of a field test. That is almost always done as a part of the natural process of program development, and it is without doubt essential. But the more fragmented process assures that the separate ingredients of an experiment may be of sufficient quality, but usually tells us little about the resulting mixture.

To be more concrete, consider what a pilot feasibility experiment may tell us about problems which can (and do) occur in major field experiments. The chronic problems, judging from precedent, bear on: the target population; the response variable; the treatment delivery; and randomization. Except for the last, difficulties with each item has surfaced in most program evaluations, randomized or otherwise.

4.1 Target Population

The chronic problem here is that members of the target population, those individuals or institutions which are supposed to avail themselves of a novel program, are *not* well identified. That is, one usually has a general idea of who might be interested, deserving, and so on, but prior to a major field experiment it's usually not at all clear how one is supposed to identify those individuals quickly, screen them, involve them in the research, and so on.

So, for example, a "need" is declared, a program developed, and field test mounted without knowing exactly who is needy and how to get at

them. The so-called career education programs were beset by this problem for at least two years before really coming to grips with it. The absence of any hard information on which adolescents were needy or even interested in career education, difficulty of setting up a good system for referring adolescents from their local high school to an as yet untested and poorly understood novel program, generated problems in assuring a decent sample size for the treatment groups much less for the control condition.

Exactly the same kind of problem occurred at about half the sites of the Section 222 experiments. These admirable tests ran straight into the problem of recruiting and selecting individuals for treatment, i.e. day care, because the size and nature of the relevant local target population was not sufficiently well known, referral services for the new program had to be set up with great effort since neither physicians nor hospital discharge officers were knowledgeable about either the new program nor the fundamental need for randomization. The problem appears to have stabilized during the first year of the experiment's conduct.

Remarks. One of the best conditions under which a randomized experiment can be established is one in which the demand for services, the number of members of the eligible and interested target population, greatly exceed the supply of treatment facilities. With a new small program, the latter condition is often met naturally. But the former condition can only be known through needs assessment surveys or through pilot tests of the kind suggested here. The market needs to be identified well before the experiment and to be expanded where necessary to enhance the feasibility of a randomized trial.

4.2 Response Variables

In the behavioral and social sciences at any rate, the character of a dependent variable, especially a newly developed test or rating system, is often insufficiently documented prior to a major experimental test. The problem is chronic and, more importantly, critical in fair estimation of program effects. In brief, the response variable's relevance to the treatment program is often quite low, despite its "face validity." And it is through research *prior to* the main experiment that the most direct evidence can be obtained, that the best systems for assuring relevance can be set up.

For example, standardized achievement tests have often been used as a response variable in appraising the impact of compensatory education programs. But, in fact, many such programs do not focus on academic achievement of deprived students even when they are supposed to do so. Even when they do, students in the needy category often perform so poorly that the test is simply insensitive to their true level of achievement and to changes in that level. To be sure, the test result may also be affected notably by local testing conditions which produce anxiety, apprehension, or confusion among students, factors which are bound to depress test scores generally. Similarly, in health-related programs, measures of (say) functional mobility of the aged or arthritic may be quite reliable when made with well trained raters. But in the field where conditions of measurement are not ideal, even the well trained may yield ratings which contain a good deal of random variation or systematic irrelevance. And if the program itself directs only a little attention to improving functional mobility, then unreliability will make the subtle effect difficult or impossible to detect.

Now aside from the normal precautions to assure reliability of measurement and relevance of the response variable, which incidentally are often not taken, a pilot feasibility test appears to be a decent approach to accommodating the problem. Prior to the main field test one obtains all the evidence one can on the sensitivity of the measures. And the test should help one understand the kinds of quality-control devices and record management tactics which should be employed in the main study to assure the integrity of the data.

4.3 Treatment Delivery

If the main field experiment directs attention to impact when the program is delivered, it is natural to focus a pilot field test on the matter of actual delivery.

That is, during the pilot test phase, the kinks in the delivery system are worked out. Mechanisms are developed to assure that an individual who is supposed to receive an income subsidy does indeed receive it and no other. A verification system is set up to assure that students who are supposed to participate in an activity do indeed do so, and so on. This basic requirement that one establish procedures for monitoring delivery seems trivial. But in fact it is not always a simple matter. The New Jersey Negative Income Tax Experiments generated grand-jury hearings when it was discovered by journalists that, unbeknownst to the experimenters, some treatment group subjects were receiving multiple subsidy payments to which they were not entitled.

A second chronic problem concerns the individual's willingness or attentiveness in receipt of treatment when assigned to the treatment condition. For example, in the Kaiser Permanente experimental tests of multiphasic screening, many of the individuals assigned to the free screening program failed to turn up for their periodic examination. The Kaiser staff, interested in the preventive benefits of screening and not in turn out rate, mounted an intensive effort to encourage participants to come for screening. The battery of telephone operators who furnished oral reminder

jacked up the participant rate to a stable 65% for the ten-year period of the experiment. A similar encouragement strategy was developed during the course of experiments to evaluate the children's television program "Sesame Street."

Here, the encouragement strategies were developed on line, i.e. during the conduct of the main experiment. It's likely that at least some problems could have been reduced earlier through pilot tests.

4.4 Randomized Assignment and Maintenance of Condition

96

The preceding section dealt with maintaining a regimen, and here we consider both that maintenance and the assignment process. The idea of the pilot test in this instance is to anticipate and accommodate problems which we expect will otherwise arise in the main test, to develop some ideas about the problem's severity, and to develop and test strategies for accommodating the problems.

The pilot test looks at the question "How can randomized assignment be accomplished best?" and proceeds to examine tactics for enhancing feasibility of randomized assignment in subsequent main field tests. So, for example, the Diet Heart Feasibility Study helped to determine if indeed randomized assignment of individuals to alternative cholesterol reducing diets was managerially possible, ethically acceptable, and socially innocuous. In a more elaborate pilot test, various public arguments for randomization might be tried out, various mechanical techniques for achieving randomization unobtrusively might be tested, and various systems for controlling the inevitable lapses in randomization might be examined.

Maintaining individuals, once assigned, in the alternative levels of treatment, or in alternative treatment regimens, or in the control condition if there is one is important, of course. And in the absence of any prior information about alternative methods of doing so effectively, a pilot test of a chosen approach seems prudent. For the treatment conditions, systematic encouragement and reminders for an effective tool and their worth or worthlessness should be evident in a pilot test. For members of a no-treatment control condition, additional incentives for participating in the experiment may be warranted (see remarks below). Those may be tangible or intangible, but in either case, their usefulness ought to be established before the main experiment is put into the field.

4.5 Summary

To summarize, the most direct way to establish the feasibility of a large field experiment is to mount a pilot field experiment. That smaller test can help one to identify unexpected problems, to try out solutions to the problems we know are chronic, and to accumulate information which is often essential to the quality of a major field test. With very novel programs whose character is not well understood by the public, whose target population is difficult to reach, whose effects may be subtle and virtually undetectable using off-the-shelf measurement devices, such a pilot test is essential.

With programs backed by intensive longer term research on target populations, on response variables, and so on, the pilot test is less crucial. It becomes considerably less crucial when the experimenter already knows a good deal about mounting very high-quality field surveys in general, and field experiments in particular.

A pilot test may itself not be practical when time is short, resources are slender, and a conservative approach is not warranted. In that case, one can only try to work out tentative solutions for some of the problems we've identified and be ready to improve them during the main experiment if they prove inadequate.

5. Direct Constraints on Feasibility of Randomized Tests

There are a variety of difficulties which can be anticipated to assess feasibility of an experiment. Both the difficulties and some tactics which can be used to resolve them are discussed in the following remarks. Since both irrelevant factors, i.e. red herrings, and pertinent factors may influence judgments about feasibility, so both kinds are discussed here.

5.1 Randomization and Selection

Basic misconceptions about randomized experiments can affect judgements about their feasibility. We consider one such misconception here in part because it emerges almost invariably in discussion with lay audiences about whether an experiment can or should be done.

The misconception concerns the idea that treatment group members be selected randomly from a prescribed population. This is often impossible, especially where individuals must volunteer for the program, and so one must reach the judgement that a randomized experiment is impossible.

Now strictly speaking, randomization in an experiment refers to the *assignment* of individuals from *a pool of eligible candidates* to program variations or alternatives. Virtually nothing about how the initial pool of candidates was actually constructed need be implied.

For example, candidates who apply for admission to a manpower training program necessarily include only those individuals who have heard about the program; many have low salaries and poor skills, which given them some incentive to apply for admission. The resultant pool of applicants will not ordinarily be representative of the

total population of people eligible for manpower training. Nonetheless, we can still conduct a legitimate experiment, randomly assigning applicants to training variations, in order to compare the relative effects of those variations. It is the random assignment process which is crucial to the unbiased estimation of relative effects on the candidates at hand. This is not to say, however, that the process of constructing the pool of candidates for an experimental test is unimportant. Indeed, it is important in that it determines how generalizable the experimental results must be. Suppose, for example, that only early applicants for a training program constituted the basic pool of candidates. After randomly assigning members of the pool to program variations, we might find that one particular variant of the program, say skill training and general education, was more effective than skill training alone in increasing job opportunities. It is quite possible that this result is not generalizable to *late* applicants to the program, although it is legitimate with respect to early applicants. Those who apply late may be delayed by their inability to read or to monitor governmental services, or for other reasons, and they may profit greatly from general education components added to their skills training. Making generalizations about the program's impact on groups not represented in the experiment can be hazardous for this and other reasons. So some experimental tests involve not only random assignment of individuals to program variants but random selections of individuals from a population of eligible candidates as well. Randomized selection, of course, is not the only determinant of generalizability in evaluations, experimental or otherwise. Others are examined briefly below.

5.2 Randomized Assignment to Control: Shifting Treatment Variations

One of the most frequently mentioned obstacles to the conduct of randomized tests concerns the random assignment of individuals to treatment or control conditions. There are at least four issues implicit to arguments about this matter, and we consider each in turn. The first is a matter of design of the evaluation and involves a shift in the question which the experiment is supposed to answer. This option is considered here, and other options which may be taken to determine or more directly enhance feasibility are discussed in the next three sections.

It is clear that in some cases, it will be illegal, unethical, or otherwise imprudent to assign some members of a target sample to a "control" (no-treatment) condition. Nonetheless, it is still possible to conduct randomized experimental tests without losing sight of the basic aim: to understand the nature of program effects. Specifically, we can compare the relative effectiveness of program *variations* using randomized tests where it is important to determine if some of those variations are more effective than others.

It may in any effect make more scientific sense to test variations. One would often like to know how response varies with different levels of intensity or elaborateness of treatment, not merely what the effect is at one level. One would often like to know whether a more expensive program or program component is that much more effective than a cheap and different program or component which is advertised to have roughly the same effect. In the latter case, the economic justification for testing variations is also clear.

To be more specific, consider a special police training program designed to reduce assaults on police. It may be funded well enough to accommodate all eligible candidates. Under this condition, policemen who are randomly allocated to a no-program (control) condition may object to their assignment and resist participating in an experiment. Managerial interest in and logistical support for a control condition may not be available for a variety of reasons, despite the fact that it is not at all clear that the program itself will be effective. To deal with these problems, it may be possible to test several program variations against one another; or to test expensive elements of the program against one another rather than to try to test the complete program against control conditions. This strategy will at least provide an unbiased estimate of the relative impact of important training variations (in reducing assaults, say). And if the experiment examines expensive program elements, we will be able to determine which of those elements are least useful in reducing assaults. *Not using a control condition forfeits the option of estimating program effects on assault relative to no program at all. But the option itself may be useless in the sense that "no program" is not a politically feasible alternative.*

The comparison of program variations need not be justified solely on grounds that control group members may feel deprived. There are important ethical reasons for using a variations design which are discussed below (see Ethical Grounds for Criticism). And there are still other cases in which comparisons among both variations and the control condition are warranted. For example, in evaluating the impact of a Manpower Development and Training Act program in Virginia, Brazziel (1967) suggested that, because the vocational program could not accommodate all eligible candidates, the candidates be randomly assigned only to program and no-program conditions. In addition, however, he did take the opportunity to develop a major program variation—general education plus vocational training—against which the regular program could be compared. Eligible candidates were then assigned to one of three conditions: vocational training, gen-

eral education plus vocational training, and a control condition. In the event of failure of the vocational program versus no-program comparison, a comparison of the program variations would still be useful to determine if the program variation (general education plus vocational training) leads to trainees who are better equipped to adapt to different job requirements than those who receive vocational education alone.

5.3 Randomization and Differential Effects of Treatment

One of the special constraints on many program evaluations is that different types of people may be in need of treatment, and effectiveness of treatment may vary with person type. Accommodating that constraint is not difficult, provided that the person type can be accurately identified. Two cases are considered below. In the first, we focus on experiments which reveal whether indeed there is an interaction between person type and treatment type. In the next section, we focus on the case in which randomization and need for what is believed to be effective treatment are at issue.

Even a cursory investigation of textbooks on experimental design reveals strategies which can be used routinely to determine how different types of people are affected differentially by a program. Given these general designs, it is up to the evaluator and the program developer to speculate on what attributes of people might interact with the program's effects and to decide upon a reliable way of discovering whether people have those attributes. The speculation may be based on anecdotal information as well as more structured judgments of the informed program developer. And if one can measure those attributes well before the experiment, they can be incorporated into a randomized block design which will permit us to detect the interaction when it occurs.

Such designs have often been used by sophisticated analysts. Results of some California Youth Authority experiments, for example, suggest that delinquent boys who are socially assertive do have the capacity to work in and benefit from confrontive group treatment, while boys sensitive to threats fare better under more supportive treatments which de-emphasize confrontive, probing behavior (Knight, 1970). At a cruder level, the Health Insurance Experiment mounted programs in different sites to assure that if effects of insurance (say) on health services utilization vary with local access to Health Maintenance Organizations, or with site-to-site differences in use of health services, the experiment will detect those interactions. Good experimental tests of clinical treatments regularly incorporate qualitative characteristics of clients into designs not only to detect differential effects of treatment but also to anticipate problems in field implementation of the program. The Cohen and Krause (1971) experiments on therapy for wives of alcoholics, for example, deliberately included demographic variables to accommodate the known tendency of clients from upper socioeconomic classes to seek and begin treatment more quickly than individuals from the lower-income brackets, to be more accessible to program staff, to be more easily engaged in treatment, and so forth.

By ignoring the possibility of such interactions, of course, we run the risk of not detecting the program's main effects. One might find, for example, that there is no difference between two programs, when in fact one program affects type A individuals dramatically in one direction while the second program afects them equally in the opposite direction. Conversely, we also run the risk of adopting a program for general use (on the basis of large average effects) when in fact the effects differ considerably, depending on characteristics of particular subgroups in the target population.

5.4 Randomization and Need for Treatment

The preceding section focused on changing the character of the treatment and the evaluation question to accommodate the problem of resistance to randomized assignment to treatment and control condition. Here the focus is also on changing the evaluation design, but alteration is made to screening tests for the target population rather than the program. The objective is the same: to avoid or attenuate a possible local constraint on randomization, and so to enhance feasibility of an experiment.

Randomization is most appropriate when the effect of the treatment variation on the sample at hand is unknown. We recognize that the effect is unknown from the judgements of experts. They regard the evidence as equivocal and, in the absence of any other information, so usually must we. Now this immediately suggests that as a general strategy in identifying the target population to which the program is most relevant, one ought to classify possible recipients into three classes: those who, most experts would agree, need the program; those whose need is debatable or ambiguous; and those who clearly do not need it at all. It is the middle group which is most pertinent to randomized assignment, there being no other rational basis for providing treatment.

The best example which we have been able to find to illustrate this perspective is the British Myocardial Infarction study, mounted to determine whether home care or hospital care is a better vehicle for treatment of a certain class of heart attack victims. The serious condition of some patients, physicians said, clearly warranted intermediate-term hospital care; for others, such care was very likely to be a waste of time. The gray

area of need included patients for whom a confident judgement could not be made, and it was members of this group who were assigned randomly to home or hospital care in the experiment. The group had until then almost invariably gone to hospital rather than home since hospitalization costs were paid, physicians had been very conservative in their judgements, and for other reasons. The experiment, carried out successfully, was useful in obtaining evidence that home treatment was no less effective than hospital, and in obtaining data useful for economic planning and management of a broadened home care system.

An experiment of this type tells one virtually nothing about the impact of the program on those who are said to be really needy. But it does do so for the ubiquitous marginal group. If the experiment is informative for this group, then the same theory might be extended to an adjacent group, said to be needy, but now constituting a new marginal group, to see if the treatment has some impact on them.

Remarks

Even with initial prior agreement by expert judges to label the marginally needy, the actual experiment may fail because the judges, on second thought, may find they can really assign very few to the marginal group.

This appears to have occurred in judicial experiments, where prior judicial agreements to label those for whom a sentence is quite arbitrary were abandoned during the course of the research. They appear to have occurred in experimental tests of parent effectiveness training where the agreement was subverted by staffs with a strong vested interest in the outcome of the experiment. And it has occurred elsewhere. The problems and potential solutions in these instances might be better identified in a pilot field test rather than in a large-scale effort.

5.5 Cost of Randomized Experiments

We often hear the claim that experiments are rather expensive and time consuming. Yet the detailed costs of most program evaluations, experimental or not, are often poorly documented, suggesting that contentions about expense cannot be easily verified. The data necessary to permit a fair comparison between, say, a randomized test and a very well thought-out and quasi-experimental test are simply unavailable. To be sure, some evaluators have laid out the costs of evaluation well (e.g., in the Taiwan Fertility Control research), but most have not. More generally, there exist no special accounting conventions for costs of program evaluation and no coherent body of statistical data on costs. The National Institute of Education, in fact, has had to develop special contracts to lay the groundwork for good accounting practices for documenting the costs of the experimental evaluation of the Career Education programs which it supports.

The only hard comparative data of which I am aware, bearing on the costs of experiments versus other methods of impact evaluation, stem from the NIE effort. Randomization appears to have required much less than a 1% increase in evaluation budgets, the increase being spent on payments to control group members and to experimental group members in return for their cooperation. The data are based on Experience-based Career Education Programs which shifted from their plans to conduct nonrandomized assessments (covariance analysis) to randomized tests of their programs.

If we examine other precedents more closely, it becomes obvious that not all experimental tests of social programs need be costly in absolute terms. Especially in education, the feasibility and utility of small, economical experimental tests of less than a year duration have been demonstrated repeatedly. For example, Goodwin and Sanders (1972) required less than three months to collect evidence on the effectiveness of tape-recorded curriculum units for use on school buses; Zener and Schnuelle's (1972) assessments of alternative career education programs for high schools took less than 12 months. The Welch and Walberg (1972) experiments on dissemination of teaching materials for Project Physics (Harvard) required less than 12 months and $30,000 to complete. Other economical experiments in evaluation of curriculum and teaching strategies are described in Riecken et al. (1973), Gage (1963), and elsewhere.

Experiments especially need not be costly if the treatment is of short duration and if the time interval between imposition of the program and the observation of the program recipient's response is small. For example, in the Manhattan Bail Bond Experiment (Botein, 1965), the program consisted of a bail waiver for individuals accused of having committed certain crimes, followed within a year by observation of a criterion variable—failure of the accused to appear for trial. Similarly, experimental evidence regarding effects of various voter registration campaigns was available soon after the new campaigns were tried (Gosnell, 1929). The effects of alternative communication strategies are available soon after the subjects' receipt of information; for example, the classic wartime propaganda and communications research of Hovland, Lunsdaine, and Sheffield (1949). In marketing and census research, information about the relative effectiveness of various methods of eliciting and transmitting valid data from respondents can be made available routinely within six months after survey programs are initiated.

This is *not* to say, however, that some experimental tests have not been expensive and time

consuming in absolute terms. Those programs which are expected to have long-term effects or to have effects only after a long period of treatment can be particularly expensive. Staff required for evaluation must be maintained, and decisions about wholesale adoption of the experimental program are delayed until data are obtained and analyzed. The Negative Income Tax Experiment is an expensive (more than $12 million) and long-term (6 years) research effort, where time is required primarily to "fix" the experimental treatment (i.e., to get people familiar with the welfare subsidy) and to determine long-term effects of the subsidy. Experimental tests of criminal reform programs, of rehabilitation strategies for the mentally ill, and of some education programs are time consuming, not only because the time necessary for treatment can be long, but because it is the long-term rather than short-term effects that are most relevant to program development.

At least with respect to absolute size of investments, the requirements of experimental tests vary considerably with the particular developmental stage of the program, the adequacy of short-term effects as an indicator of program success, and the time necessary for completing the treatment program. There are at least two important issues, however, which suggest that we cannot be content with decision of absolute costs: the costs and benefits of lower quality appraisals, and the intermediate products of experimental evaluations. The cost of not doing an experiment will often be high, simply because the data stemming from observational studies will usually be equivocal, and the cost of wrong decisions (or no decisions) based on equivocal data can be high. Unfortunately, there have been few formal analyses of the costs to society of not doing evaluations, of doing equivocal evaluations, or of mounting rigorous tests of social programs. The better (and perhaps the only) cost/benefit analyses of experiments are in the fertility-control area where, for example, the Population Council has succeeded in obtaining fairly good information on the cost and impact of data stemming from its fertility-control research.

On the other hand, there has been a bit more progress in identifying the benefits of evaluation and of staging research to obtain usable products periodically before the experiment's completion. Before program effects appear, the experiment often provides better information about the program's target group than was previously available. Such baseline data often yield more accurate characterizations of the target group than were available at program inception, and consequently may be helpful in designing and launching subsequent programs. See Field and Orr (1975) for remarks on this in the context of the Housing Allowance Experiments and the Negative Income Tax Experiments.

5.6 Accommodating Ethical Constraints on an Experiment

Claims about the ethical aspects of randomization generally take several related forms. The contention that control (no-program) group members are deprived of a program which might be beneficial to them occurs often. A mirror image of this complaint is that the program recipient is subjected to risk by his participation because a novel program may have unpredictable negative effects. A second broad class of criticisms concern manipulation of human beings—an activity which may be objectionable in principle. A related issue concerns the notion that the research subject is being exploited regardless of the costs and benefits of the experiment; that is, that he receives little in the way of direct reward for his participation and lacks even a guarantee that the information he provides will not be used improperly.

Some experiments can be judged to be unethical for these reasons. But this does not imply that all experiments are unethical, any more than one high-quality experiment implies that all are of high quality. The following remarks capitalize on what we already know about fairly universal if crude ethical standards and about potential conflicts between those standards and experimentation. They focus on the question of how to design the experiment within the framework set by good ethical standards.

Failure to experiment as unethical. A frequent claim about randomized experiments is that some members of the social program's target population—the control group members—must be deprived (randomly) of a benefit. The claim assumes, of course, that the treatment is actually beneficial, and if it is *known* to be beneficial, then the experiment may well be unethical. But the aim of most experiments is to discover whether there is a detectable program effect; we may not need an experiment at all if the impact is already understood. By restricting randomization to programs about which we are in doubt, we avoid the ethical dilemma (or accusation) of depriving an individual of a benefit. There can be no benefit if the program is useless and often we cannot show if it is useful without an experiment.

A related line of argument here is that a failure to discover whether a program is effective is unethical. That is, if one relies solely on nonrandomized assessments to make judgments about the efficacy of a program, subsequent decisions may be entirely inappropriate. Insofar as a failure to obtain unequivocal data on effects leads to decisions which are wrong and ultimately damaging, that failure may violate good standards of both social and professional ethics (Rutstein, 1969). Even if the decisions are "correct" in the sense of coinciding with those one might make based on randomized experimental data, ethical problems per-

sist. The right action taken for the wrong reasons is not especially attractive if we are to learn anything about how to effectively handle the child abuser, the chronically ill, the poorly trained, and so forth.

Design of Ethical Experiments

There will always be cases in which the use of a no-program control condition conforms readily with professional ethics. That is, there is agreement that program effectiveness is ambiguous, that the available data are insufficient for making a judgment about its quality, and hence an experiment is ethically justified. But a public or standard may deviate from this notably, and it may become necessary to adopt some strategy for either altering that public ethic or adjusting the design to accommodate it.

Changing a public ethic is usually impossible with the time available to mount experiments. Nonetheless, some short-term approaches have been tested. Some rely heavily on the use of the media to enhance the reading public's understanding of the process. That some journalists and science writers can effectively translate the matter into lay terms is readily evident from articles by Alan Otter in the *Wall Street Journal*, Koutalek in the *Chicago Tribune*, P.C. Gilmore in the *New York Times*, and elsewhere. Alice Rivlin has written special articles on the Negative Income Tax Experiment for the *Washington Post* and *New York Times*, as other social scientists have done for other presses.

More direct action is usually warranted, including the construction of unobtrusive but effective schemes for randomization, and fair sets of instructions to assure that informed consent requirements for participants are met. This area does not seem to have received much in the way of systematic research and development. The whole matter of encouraging participation in an experiment is still a very ill-documented area. The little systematic research we've seen suggests that people will find randomization more palatable if they are party to the randomization process: they pick the lottery number themselves rather than having someone else do it. They will find it more palatable if even being a member of a control group affords some benefit (see Section 5. and remarks below). They will find it more palatable if there are intangible benefits, such as increased self-esteem or decreased anxiety or loneliness or boredom, by participating.

There will also be cases in which a randomized test of a program versus a no-program control condition is unethical. That a particular experimental design is ethically unacceptable *in a particular evaluation* of course, implies nothing about the acceptability of other randomized designs. In fact, a variety of techniques have been developed to reduce or eliminate conflicts between ethical standards and evaluation needs.

One obvious device is to stage the introduction of treatment so that one merely *delays* treatment for individuals in the randomized control group. The strategy is sometimes essential in any event because many programs cannot accommodate all eligible candidates immediately, and staged acceptance of candidates is managerially justified. The control groups may subsequently be reduced incrementally or all at once, so long as the delay is sufficient to permit useful comparisons between program participants and nonparticipants. (See Chapter IV of Riecken et al., 1974, for more detailed description of this design and its limitations.)

"Playing the winner" is a related strategy, used more often in bio-medical research, to estimate program effects with minimal deprivation to members of the less effectively treated group. Here, subgroups of candidates or individuals are assigned to a program only as long as the outcome of treatment is successful. When a failure occurs, the very next subgroup or individual is assigned to the control (or alternative treatment) condition. Subgroups continue to be assigned to the control group so long as no failure occurs. When it does, the very next subgroup is assigned to the first treatment. And so on. This strategy is a more complex one, but recent analytic work shows that it can be very effective when success or failure become evident quickly and when switches can be accomplished easily (see, e.g., Fushimi, 1973). The strategy also requires that the "success" be readily identified when it occurs, a demand which may be difficult though not impossible to meet in some settings.

If delays in program participation are ethically unacceptable and if program installation involves no naturally occurring delays, then other strategies can be used. Rather than think solely in terms of treated versus untreated program candidates, for example, it is often reasonable to change the research question slightly to permit us to think about comparing treatment variations, an option already discussed in Section 5.2 above. Candidates for the social program can be allocated randomly to different levels of treatment, the lowest level being a minimal ethically acceptable offering. This idea has been used in both the Negative Income Tax Experiment and Health Insurance Experiment, where deprivation of economic benefits relative to current social standards would be ethically unconscionable despite its importance as an economic question. And it has been used in critical medical studies such as Rutstein's (1969) tests of cortisone against aspirin in treatment of rheumatic fever. In these and other cases, new programs are compared against the better conventional ones rather than against no program at all in order to satisfy both scientific and ethical standards.

Similarly, experimental assessments of *components* of a program rather than the total program may also be possible when there is little prior evidence on effects of the components, but there is strong professional or societal belief that the program is indeed effective. Physicians, for example, are often confident that integrated health-care systems are good and that if "total" health care is delivered to an individual, his health will improve. Under these conditions, it may be impossible to mount a fair test of the total program. Instead, the evaluator might look for those components of the program about which there is some doubt as to their effectiveness. For example, integrated health-care delivery systems in lesser developed countries are being supported by the U.S. Agency for International Development. To the extent that nutrition, health care information, and the like are regarded a priori as "a good thing," trying to evaluate their total effect using a randomized experiment may be a pointless exercise at this time. Component-wise evaluation is not. No one knows, for example, how paramedics should be chosen (monks, midwives, relatives, or village elders) and trained to yield high treatment rates with minimal cultural disruption. The situation presents us with an opportunity to experiment with alternative recruitment and training strategies even if we do not obtain unequivocal data on the actual product delivered by the trainees.

Often, criteria such as merit or need are justified on ethical grounds for assigning individuals to programs whose effects are not well documented. And the meritocratic criteria lead some critics to conclude that randomization is therefore impossible on ethical as well as managerial grounds. However, we can still obtain evidence based on randomized tests if we capitalize on so-called regression-discontinuity designs (Thistlethwaite & Campbell, 1960). In the simplest case, one orders all program candidates on the basis of need, then assigns all obviously deserving candidates to the program and all the obviously undeserving to the control condition. Individuals in the ubiquitous marginal group are assigned *randomly* to program and control conditions; their marginality implies that no reliable judgement can be made about the extent to which they merit the program. A variant on this design has been used successfully in the British myocardial infarction studies, where marginally ill individuals were randomly assigned to home or to hospital care to satisfy ethical standards and to discover whether hospital care resulted in any notable improvements in their health.

Demands on the research participant. One of the simplest ethics-based criticisms of randomized experiments is that regardless of the scientific and social benefits of the experiment, it is a distinct imposition on the research participant. Exactly the same criticism, of course, can be leveled against survey research of any sort and against quasi-experimental and other types of evaluative research. The research participant does indeed provide a service to the researcher—information about himself, his time, energy, and courtesy in providing the information, and so forth. And insofar as the social scientist profits (at least intellectually) from the information he receives, why should not the provider also profit? The rewards to be sure need not always be large or even tangible. For example, there is some evidence for the contention that in certain types of research, the interviewer's behavior, conversation, and discussion of research do constitute a temporarily rewarding experience for interviewees. If higher demands are made of research participants, they may be entitled to more tangible rewards for their cooperation. Students who participate in experimental tests of NIE-supported Career Education Programs, for instance, are paid for providing their opinions, reactions, for taking tests, etc., regardless of whether they were assigned to the experimental program or to a control condition (the conventional high-school programs in career education). What the nature of the reward should be in different types of experiments and how alternative rewards affect the integrity of the experiment need more empirical investigation, however. The little data available on this topic stem primarily from survey research where alternative rewards have often been tested experimentally to determine how rewards such as money payments, small gifts, etc. stimulate cooperation (e.g., the Sudman & Ferber (1971) work on strategies for improving response rate in consumer surveys).

Confidentiality of information. The problem of assuring confidentiality of data is not confined to experimental research but appears in survey research as well. But the problem has been highlighted by the Negative Income Tax Experiment, in which a county prosecutor forced economic researchers to yield research records on identified subsidy recipients (research subjects). The case is a regrettable illustration that the researcher may be cast unwillingly into the role of informant, if he does not anticipate the possibility of judicial or legislative appropriation of his records for prosecuting some of his research subjects. There have been some advances in resolving this and related conflicts. For example, procedural and statistical devices have been created to assure confidentiality of respondents' reports without undermining research goals (Boruch, 1974). Special forms of testimonial privilege for social researchers are being constructed to supplant or complement technical devices for assuring that research records are used only for research purposes (see Reicken et al., 1974; Boruch, 1976, and references therein). These approaches are imperfect, but they are being field tested, and they do help to reduce conflict between legal demands for individual records

and the social scientist's ethical requirement for confidentiality of record on his respondents.

References

Bernstein, I., & Freeman, H.E. *Academic and entrepreneurial research.* New York: Russell Sage, 1975.

Boruch, R.F. *Assuring confidentiality in social research.* Book manuscript in preparation. Evanston, IL.: Northwestern University, Psychology Department, 1976.

Boruch, R.F. Bibliography: Illustrative randomized field experiments for program planning and evaluation. *Evaluation,* 1974, *2,* 83–87.

Boruch, R.F. On common contentions about randomized field experiments. In R.F. Boruch and H.W. Riecken (Eds.), *Experimental testing of public policy.* Boulder, Colorado: Westview Press, 1975. Pp. 107–145.

Boruch, R.F., Magidson, J., & Davis, S. *Interim report: Secondary analysis of Project Middlestart.* Paper presented at the annual meetings of the American Psychological Association, September 1975.

Boruch, R.F., Wortman, P.M., & DeGracie, J.S. *Executive summary: Project on Secondary Analysis.* Presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April 2, 1974.

Botein, B. The Manhattan Bail Bond Experiment. *Texas Law Review,* 1965, *43,* 319–331.

Brazziel, W.F. Effects of general education in manpower programs. *Journal of Human Resources,* 1966, *1,* 39–44.

Campbell, D.T., & Boruch, R.F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C.A. Bennett and A. Lumsdaine (Eds.), *Central issues in social program evaluation.* New York: Academic Press, 1975.

Campbell, D.T., & Stanley, J.C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Clarke, R. V. G., & Cornish, D.B. *The controlled trial in institutional research.* London: Home Office Research Studies, 1972.

Cohen, P.C., & Krause, M.S. (Eds.). *Casework with wives of alcoholics.* New York: Family Services Association of America, 1971.

Conner, R.F. *A methodological analysis of 12 true experimental program evaluations.* Ph.D. dissertation, Psychology Department, Northwestern University, Evanston, Illinois, 1974.

Cook, T.D., & Campbell, D.T. The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1975. Pp. 223–324.

Daniels, D.N., et al. *DANN services program* (Research report, National Institute of Mental Health, Grant No. 02332). January 1968.

Deniston, O.L., & Rosenstock, I.M. *The validity of designs for evaluating health services.* Research report, Ann Arbor: School of Public Health, University of Michigan, March 1972.

Field, C.G., & Orr, L.L. Organizations for social experimentation. In R.F. Boruch and H.W. Riecken (Eds.), *Experimental testing of public policy: The proceedings of the 1974 Social Science Research Council Conference on Social Experiments.* Boulder, Colorado: Westview Press, 1975.

Fushimi, M. An improved version of the Sobel-Weiss play-the-winner procedure for selecting the better of two binomial populations. *Biometrika,* 1973, *60*(3), 517–523.

Gage, N.L. (Ed.). *Handbook of research on teaching.* Chicago: Rand-McNally, 1963.

Goodwin, W.L., & Sanders, J.R. *The use of experimental and quasi-experimental designs in educational evaluation.* Research report. Boulder: Laboratory of Educational Research, University of Colorado, 1972.

Gosnell, H.F. *Getting out the vote.* Chicago: University of Chicago Press, 1927.

Heber, R., et al. *Rehabilitation of families at risk for mental retardation.* Madison: University of Wisconsin Rehabilitation Research and Training Center, 1972.

Hendricks, M., & Wortman, C. *Reactions to random assignment in an ameliorative social program as a function of awareness of what others are receiving and of outcome.* Evanston, Illinois: Psychology Department, Northwestern University, 1975.

Hill, A.B., Marshall, J., & Shaw, D.A. A controlled clinical trial of long-term anticoagulant therapy in cerebrovascular disease. *Quarterly Journal of Medicine,* 1960, *29,* 597–609.

Hold, N. Rational risk taking: Some alternatives to traditional correction programs. *Proceedings: Second National Workshop on Corrections and Parole Administration.* San Antonio, March 1974. College Park, Maryland: American Correctional Association, 1974.

Hornik, R.C., Ingle, H.T., Mayo, J.K., McAnany, E.G., & Schramm, W. *Television and educational reform in El Salvador* (Research Report No. 14). Stanford, California: Institute for Communications Research, Stanford University, August 1973.

Hovland, C.I., Lumsdaine, A.A., & Sheffield, F.D. *Experiments on mass education.* Princeton, N.J.: Princeton University Press, 1949.

Knight, D. *The Marshall Program: Assessment of a short term institutional treatment program, Part II: Amenability to confrontive peer group treatment.* Sacramento: California Youth Authority, 1970.

McKay, H., McKay, A., & Sinesterra, L. *Stimulation of intellectual and social competence of Colombian preschool-age children affected by the multiple deprivations of depressed urban environments* (Second Progress Report). Cali, Colombia: Universidad del Valle, Human Ecology Research Station, September 1973.

Meier, P. The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J.M. Tanur, F. Mosteller, W.B. Kurskal, R.F. Link, R.S. Pieters, and G. Rising (Eds.), *Statistics: A guide to the unknown.* San Francisco: Holden-Day, 1972.

Porter, A.C. *Introduction to NIE's Program on Measurement and Methodology.* Washington, D.C.: National Institute of Education, March 1, 1975.

Riecken, H.W., Boruch, R.F., Campbell, D.T., Caplan, N., Glennan, T.K., Pratt, J.W., Rees, A., & Williams, W. *Social experiments: A method for planning and evaluating social programs.* New York: Seminar Press, 1974.

Ruffen, J.N., Grizzle, J.E., Hightower, N.C., McHardy, G., Schull, H., & Krisher, J.B. A cooperative double-blind evaluation of gastric "freezing" in the treatment of duodenal ulcer. *New England Journal of Medicine,* 1969, *281,* 16–19.

Rutstein, D.D. The ethical design of human experiments. *Daedalus,* 1969, *98*(2), 523–541.

Sackett, D.L. *End results analyses in a randomised trial of nurse practitioners.* Research memorandum. Hamilton, Ontario: McMaster University Medical Center, Burlington Study Group, 1973.

Sudman, S., & Feber, R. Experiments in obtaining consumer expenditures by diary methods. *Journal of American Statistical Association,* 1971, *66,* 725–735.

Thistlethwaite, D.C., & Campbell, D.T. Regression-discontinuity analysis. *Journal of Educational Psychology,* 1960, *51,* 309–317.

Wargo, M.J., Campeau, P.L., & Tallmadge, G.K., with assistance of Lauritz, B.M., Morris, S.J., & Youngquist, L.V. *Further examination of exemplary programs for educating disadvantaged children.* Palo Alto: American Institute of Research, July 1971.

Welch, W.W., & Walberg, H.J. Pretest effects in curriculum evaluation. *American Educational Research Journal,* 1970, *6,* 605–614.

Yinger, J.M., Ikeda, K., & Laycock, F. *Middlestart: Supportive intervention for higher education among students of disadvantaged backgrounds* (Final Report to U.S. Office of Education, Project No. 5–0703, Grant OE 6–10–255). Oberlin, Ohio: Oberlin College, Sociology Department, November, 1970.

Zener, T.B., & Schneulle, L. *An evaluation of self-directed search: A guide to educational and vocational planning* (Research Report No. 124). Baltimore, Md.: Johns Hopkins University, Center for the Study of the Organization of Schools, 1972.

Development of Staff for Evaluations
(A Retrospective View)

George L. Kelling
Evaluator
Police Foundation
Washington, D.C.

George Kelling is on the staff of the Police Foundation and has been working as an evaluator in Kansas City and Dallas over the past several years. In particular, Kelling was the Director of Research for the major study of police patrol practices carried out in Kansas City. In gearing up for that project he had to put together from scratch and manage a large and complex research team. This paper presents his views on the problems that are likely to be encountered in putting together an evaluation research staff and on approaches to solving those problems.

When in confirmation class as an early adolescent, I, as many other young Lutherans, was forced to memorize Luther's explanation of the three sections of the Apostle's Creed. While no longer able to pull the explanations back into consciousness, I can clearly recall the last sentence of each explanation. The phrase, identical in each, was "This is most certainly true." The matters Luther was dealing with were, of course, eternal verities. While they may or may not be "most certainly true" for others, they were for Luther and he emphasized their importance to himself and his followers with his declaration.

As a result of administering many evaluations, I have been asked to talk to you about developing personnel for work in evaluative research. While the positions I take in the following pages certainly do not approach, for me at least, the state of eternal verities, they do achieve the level of pragmatic and survival verities in the conduct of evaluations. Part of this feeling comes from a set of values and assumptions which I have and which perhaps is worthwile for me to identify. These include:

1. It is good to complete evaluations—few really are.
2. It is good to maintain "experienced leadership" in an evaluation staff—read that: "I want to survive."
3. It is good to maintain "experienced leadership" in the organizations in which evaluations are conducted—need I explain the worth of that to you?
4. It is not that the *best* predictor of an individual's or organization's performance is his/her/its past performance—it is the only predictor.
5. And finally, conflict in the activities of organizations and personnel need not be deletorious to achievement but rather, if the rules of conflict are established, can contribute to creative and original work.

With those values, assumptions, confessions out of the way, I will continue with one final venture into the rarified air of theology with a paraphrase of a statement by Paul Tillich.

> I shall proceed to lecture now, and continue to perform in evaluations on the assumption that I am absolutely correct in what I am about to say. I am aware that I may be wrong, but I will not let that awareness interfere with this discussion of my future performance as an administrator of evaluations.

If any of you, as you read or hear this, feel like standing, applauding, and cheering, I, of course, invite you to. If on the other hand you feel like booing and hissing, there is nothing I can do to stop you, so feel free.

Verity #1. Where one's tenure is, is where one's heart is—or—*the use of consultants.*

The use of consultants is standard in evaluations and evaluation proposals. Generally consultants are luminaries from academia who have a superb record of research and thinking about methodology and/or service delivery in a particular endeavor. They are generally competent, leaders in the field, and involved in a myriad of enterprises. Generally they are capable of, and have executed, good research and/or evaluations. They are experts. Basically, they can serve two functions in an evaluation:

1. They can help "young comers" get grants, contracts and exposure. If done responsibly, this is legitimate and ought not to be sneered at. The function of a mentor or sponsor is an important one in academia. "Young comers" present a high risk to both program and evaluation administrators and grantors, but at the same time they have the energy, and are enough "on the make" to complete an evaluation. The "baptism" of "young comers" by luminaries must be un-

derstood for what it is however. Do not expect the "heavies" to conduct the evaluation or write the results. They cannot and will not.

2. They can provide technical consultation on critical points of an evaluation. Three critical points stand out:
   a. "Now that I have all this data, why did I collect it in the first place and what should I do with it." In other words, it is possible that the evaluator will get so immersed in details that he/she will forget what the original goals of the project were and how the data deals with those goals. Further, after being removed from the world of academia during the year or two of the evaluation, the evaluator may need some assistance in updating his/her statistical skills. The consultant or consultants can help the field staff of an evaluation to review their work and update skills.
   b. Review the outline for the presentation of the findings. This is related to "a" and is part of "a" yet is so important that I separate it out. Getting a good outline of the final report is *the* critical issue in getting the evaluator to put his/her pen to paper. It nicely makes a completely unmanageable task (completing the report) into a manageable one.
   c. Finally—reading the preliminary drafts of the evaluation and providing constructive, non-threatening advice. Generally upon completing the first draft, the evaluator thinks (hopes) that he/she is finished writing. In fact, he/she has just begun. Remember, *any* first draft, regardless of its weaknesses, is good. If an evaluator is reasonably good and has good consultation, *any* first draft almost assures completion.

So much for the positive contributions of consultants. They can make real and substantial contributions, but for all parties involved in the conduct of an evaluation, it is certainly best to underestimate their contributions rather than to overestimate them.

They *cannot:*

1. Supervise staff. Young, energetic staff need constant and ongoing stroking, direction, love and supervision. Consultants cannot provide that. They do not have the time, nor do they control the means and rewards necessary to manage staff.
2. Develop evaluation instruments (questionnaires, etc.). Instruments must be developed by resident evaluation staff in close collaboration with agency program staff. Consultants don't have the time, energy and, generally, the patience to collaborate as closely as necessary.
3. Write-up results. The writing of the final report is a consuming full-time task. Consultants are involved in too many things to be expected to write-up a final report.

The key thing to remember in dealing with consultants (and I do not mean this critically) is that they are *un* responsible. They are bright, knowledgeable, clever, but they have no responsibility for the final product and rarely, if ever, will be cornered into accepting such responsibility. They have different responsibilities and will meet those first—and—that is to be expected. Neither the program evaluator or administration should be surprised by this as likely they, too, are consultants some place. *This is most certainly true.*

Verity #2. The children shall lead you (or at least they will do most of the necessary "grunt work"). *Staff Structure*

I will divide this section into two parts: first, the characteristics of evaluation staff, and, second, the characteristics of the evaluation director.

Perhaps it will be easiest if I begin with the characteristics of the staff who are "on site," and who do the daily work of evaluation. (Be clear that I did not always know these verities, and not even when I knew them did I always follow them. One result is that in the early projects I have administered the casualty rate of project staff was very high. In the early days, I often took those persons for staff who were available at the time. Some were less than satisfactory. Applicants were few. I had no track record as an evaluator. Evaluation was considered inappropriate—read "inappropriate" as "sinful"—by major professors for their good students. But, I am getting ahead of myself.) The people who actually "do the daily work" of an evaluation have to have certain characteristics. These include: high levels of energy, methodological sophistication, skill at handling data, keen intelligence and curiosity, being professionally "on the make," the capability of using creatively the great freedom that evaluators have, and the ego strength to move with some comfort into an alien environment. The staff need not have, and, if you recruit the proper persons, probably will not have, organizational "smarts," familiarity with the field of service delivery, or experience in completing a project. (I will discuss these points somewhat later.)

Where are these kinds of people found? (The people who do the daily work.) The answer is quite clear. In the doctoral programs of universities. And *as* important, in the doctoral programs of *good* universities. Their characteristics are as follows:

—They have been born, bred, and expect to die in universities.

—They have never held a job (except maybe Vista or a summer camp).

—They have managed to make avoiding deadlines a fine art and skill.

—They are arrogant. (Often they *are* right—they are more methodologically skillful than their professors and, later, than you as project director.)

—They know how to develop sophisticated questionnaires but they do not know how to talk to people (read "talk" as interview without a pretested questionnaire). They will have to be driven, almost with whips, to work closely with agency program staff and to really talk to them (but once they do, another problem—that of cooptation—rears its head, which we shall discuss in detail later).

—They view all researchers and grantsmen who operate outside of universities as whores and "operators" interested only in the "bucks" (they really believe that their professors live on their salaries alone) and that all truth is to be discovered in the world by conducting methodologically "pure" experiments on freshmen.

And finally, they're marvelous. They believe the world can and will change, they work night and day, they're damned smart, and they have that marvelous characteristic of youth—energy. (Oh, I know its unbounded and undisciplined, but evaluation directors have to do something after all.)

But now in a somewhat more serious vein, I wish to talk about each of the characteristics that I find necessary in staff.

High Levels of Energy

Evaluations are difficult and time consuming. They combine all the intellectual and methodological rigors of laboratory experiments with the messiness and complications of the real world. The real world presents a myriad of problems for which a great deal of energy is necessary to solve. The following are examples.

Agency records were not devised for research. Often when computerized, they contain errors and omissions which, while not a problem for agency administrative purposes, are in such a condition that it is necessary to return to the original documents when they are used for research or evaluation.* (I don't mean to offend agency officials at this point, and maybe it is different in the medical field, but for the most part all agency data have to be verified for research purposes and every evaluation which is based upon agency data which have not been verified in great detail is a terribly suspect evaluation.)

As Mr. Lewis points out in his paper, oftentimes agency program managers who are responsible for the administration of an experiment care less about maintaining the controls of the experiment than they do about "starts" or exporting the program to other areas or jurisdictions. I would underline Mr. Lewis' point about "starts" and recommend that each of you re-read it. The dynamics and consequences of it are substantial. Given the media's interest in "starts" and the fact that *everyone* gets bored with continuing programs, the evaluator must attempt to carefully deal with and exploit both the initial publicity from "starts" and the subsequent obscurity when the experiment or program is ongoing. The management of the momentum of an experiment is critical and a balance has to be developed between the extremes of the publicity and momentum of the "start" and later obscurity of slowdown. Obscurity both has its benefits and problems. Generally, the momentum gained from the initial thrust will not provide enough energy to complete the task. Occasionally ""boosters" from agency program and evaluation staff are absolutely necessary to obtain the goal of a completed program. Alertness of the maintenance of the ongoing program is essential for evaluation staff.

A variation of the problem is "restarts." That is when an agency administrator decides that the indicator of his/her wisdom and skill is his/her ability to replicate the program in other departments, divisions, etc., before the evaluation is completed. This not only consumes a great deal of staff energy (both of agency staff who are pushing to do it and of evaluation staff who are trying to stop it) but also potentially destroys the experiment or evaluation by contaminating control areas.

Personnel involved in program efforts may have more of a vested interest in the success or failure of a program than in the conduct of the experiment and as a result inadvertently (or purposely) attempt to bias the outcome. Evaluators must constantly monitor, in as discreet a manner as possible (as monitoring itself may develop resistances) all planned stimuli, controls, and data collection.

Dealing with these and a myriad of the problems simply requires a high level of alertness and effort for a prolonged period of time. There is much "dirty work" which has to be done and on-site personnel have to have the endurance to do it. (In one city the "dirty work" meant night work for at least six weeks in a record division. That was in addition to the regular day activities.)

Methodological Sophistication

Often the exigencies of real world agency existence are such that program evaluation can be quite complicated. Finding the right design—that is an evaluation design which is as powerful as the

* Not only is a high level of energy necessary but also dealing with these sorts of problems requires a gift for great attention to detail and a toleration for the tedious—characteristics sometimes different from and in conflict with high energy levels.

program allows and warrants—requires considerable methodological sophistication. The "matching" of program and evaluation design is not to be accomplished by returning one more time to the bible of Campbell and Stanley but rather comes through the careful "wedding" of research techniques and operating programs. There is nothing mysterious about this. The evaluator simply must "muck around" in the program, data, and funds and find a design which is appropriate to the program operation, the funds available, the importance of the program, and the available data. That means that the staff must *know* design and scientific method and not just have a shopping basket of designs, one of which she/he pulls out for this program.

Skill at Handling Data

Two important things have to be said about this.

One staff member has to approach the psychological state of being an obsessive compulsive. If someone does not keep careful record of *every* decision made regarding design and data storage, the disaster of having to reconstruct those decisions will result in the waste of spending the time re-doing things and also of not meeting deadlines. Not that things cannot be reconstructed, and generally they can, but to have no way of identifying which questions were related to what indicators means a period of reconstruction beyond that normally required to re-familiarize oneself with the material. Two examples. In the Kansas City Preventive Patrol Experiment, the details and records of the sampling procedures for the community survey were never gathered together in one file or written up when the sample was drawn. When, 18 months later, we had to discuss the sampling procedures, at least three people in three different organizations had to search their files for the various memos, instructions, etc. It was possible, but that which was easy to do at one time, became complicated at another. On the other hand, in Dallas we did two departmentwide surveys. The T_1 survey was completed in 1973, the T_2 survey in 1976. Because we had carefully documented the source of every question, all coding decisions, and every other decision, the time necessary for review was spent relating the theories under which we operated to the forms of analyses we were to use. Thus, an axiom emerges. Never, never, never rely on memory. Rely on it only to fail, and, even worse, to deceive.

The second area of the importance of data handling has to do with the assessment of agency records. This is no simple matter, especially in police agencies, but I suspect in other agencies as well. Again I want to emphasize that I imply no criticism of agency records. I simply have no way of knowing whether they are adequate for administrative purposes. I assume they are. You are in a better position to know that than I. I do know, however, that almost all records will need considerable work to be suitable for research purposes. If the records are computerized, considerable work will have to be done to insure its accuracy and reliability. (Even at that evaluators must approach them cautiously since much of it is self-reported information, i.e., crime and activity analyses, which are subject to manipulation, whether conscious or unconscious, to show desired or self-serving results.) If records are kept in manual files, other problems, such as coding, or agency policies which allow for several file systems, emerge. (In one police department complaints against police officers are kept in three different places—depending on where the citizen first filed his complaint—and may or may not be stored with the other units. Notice the phrase "may or may not" since that complicates things considerably. If any officer has complaints filed against him in more than one location, and many do, the evaluator has to carefully read each one to determine if they are separate or the same complaint. Thus, even establishing the "n" of complaints is not a counting task but an analytical task.)

The evaluation staff has to know what they know, both in terms of recalling decisions and assessing data. Both tasks are far more complicated than generally thought.

Keen Intelligence and Curiosity

In some respects this is self-explanatory. But while keen intelligence and curiosity are necessary, they are not sufficient. They have to be combined with many of the other characteristics described in this section. Without energy, discipline, and creativity, intelligence simply is not enough.

Let me add one thing about curiosity as I think it to be quite important. The characteristic of asking "why" is absolutely essential. In the first place it helps to keep the intelligent person from seeing the emperor's clothes. The "emperor" can be the agency, the evaluation director, or colleagues. Secondly, it helps the evaluator pursue unanticipated findings. And, if properly pursued, these unanticipated findings can be quite important to an evaluation. It might mean the evaluation is on to something new (I call your attention to Mr. Bieck's study of police response time. The surprise finding of the length of time it takes citizens to report even serious crimes is not only of great research and program interest, but is also an indicator of just how poorly thought through the whole business of the importance of police response time has been by police, researchers, and evaluators.) or reflects an artifact of improperly stored or analyzed data. The evaluator, who, out of his/her curiosity, continues to pursue those leads, either enriches the evaluation immensely or saves it from spurious findings.

Professionally "On the Make"

Perhaps it is purely a personal matter on my part, but I simply have an easier time dealing with people who know what they want. I find it difficult to deal with people, on a project level at least, who are indecisive about their own goals. (By that I do not mean that everyone who comes onto an evaluation staff has to know that he wants to do evaluation research in a particular service delivery system for the rest of his life.) She/he may want to gain research experience, get publications, examine a service system, or do a variety of other things, but they have some sense of their own goals. If that "purposefulness" is not presented in staff members, I have been unable to develop it. (And I don't mean that an individual's goals can't change, but purposefulness remains.) The casualty rate of those who have not been purposeful has been very high.

Those people who are beginning their careers and are purposeful clearly do not yet know the prices of long hours and crash production schedules which they will have to pay to obtain what they want. But they will learn that, and most people "on the make" are willing to pay those prices. People who are not aggressively purposeful simply aren't motivated enough to "pay the price." (That makes sense. If you don't know what you want, why should you "pay the price"?)

A side comment here. People who do evaluations live on grants. Try to think of evaluation bureaucracies that do not live on grants. Few come to mind. While medical, social, police, and other service systems have ongoing existences independent of most specific projects, evaluation people either live from grant to grant, or work full-time in a university or consulting firm and do evaluations part-time. The result is that for an evaluation capacity to survive not only must it do the evaluations at hand but it must also use resources (primarily time) to generate new proposals. The alternative is constant "gearing up" or "dismantling" a staff, either one of which destroys established organizational skills and working relationships. Thus, in my judgment, evaluators must be prepared to "pay the price" of constant pressure to both complete and generate activities simultaneously. ("Workaholics" make good evaluators.)

One final comment about being "on the make." I believe that most good evaluators are from universities and will and ought to return to universities for rest and recreation (in the finest sense of recreation, that is re-creation of knowledge and skills). In order to do that, publishing is an absolute necessity. Thus, from the beginning, I have tried to insure that the data collected will be not only necessary for evaluation, but also, whenever possible, be useful as sociology, political science, or psychology and thus result in publications independent of the evaluation. I must also confess that I have not always been completely candid about this to agencies or the Police Foundation. We have called these interests the "oh by the ways." To insure the protection of agencies, I have always assured them, and *meant* it, that *nothing* will be published without their review and permission. The resultant problems are different than one would expect. First, the agencies encourage publications—administrators have found that agency reputation is enhanced by such activities. Second, and this gets to be a problem, oftentimes agency administrators get to be more interested in the "oh by the ways" than in the evaluations. (The consequence of this is that staff time can be diverted away from evaluation-specific activities to less critical issues at the *wrong* time.)

But the point is that the data, if properly collected, can be available for publication independent of whether the program succeeds, fails, never gets off the ground, or collapses in the middle (that does happen, unfortunately much, much too often) and young staff can get the publications necessary for their own careers. And, I would add, data is just too expensive to collect to be used for only one purpose. If, at no or relatively little expense, data can be collected which is multipurpose, it seems to me only prudent to do so.

Capable of Using Freedom

For some young researchers, the freedom provided in evaluation is such a burden that they just can't handle it. They search for day to day direction, are terrified of making mistakes, withdraw into obsession about codes or analyses, can't start to write a report because all they can think about is the final product rather than just the page they are on, get preoccupied with the administrative issues of evaluation rather than evaluation itself, etc., etc. At worst they begin to "rip-off" freedom, using their time for activities other than their evaluation work. (Not that I feel that staff should not be involved in other consulting, lecturing, etc., activities. I think they should. It gives them wider exposure at somebody else's expense, they enhance the reputation of the entire capacity, and it keeps them from being too narrowly focused on particular projects. But they must do so at their own expense, not at the expense of the evaluation.) Finally, they may become so cynical that termination is inevitable. They are not necessarily "bad" people; it's just that the available freedom simply leaves them unable to function.

For others, freedom is an opportunity to respond flexibly to the myriad of complexities that occur during the process of an evaluation. They (and I think I am covering somewhat similar ground as I did when I talked of purposefulness) feel comfortable making decisions and making mistakes. They are far more comfortable communicating to the project director what they have done, why they have done it, and—at first I found

this surprising—what mistakes they have made. It turns out that, while they obsess less, they are far more thorough in recording their decisions. And finally, when in a jam, they look for help. Those who can really handle freedom are open and communicative. Those who cannot, turn secretive. And once the vicious cycle of secretiveness begins, I have not yet found a way to interfere with it.

One final comment; there are good people who, at times, seem to go into a work moratorium. Generally, those periods occur during the quiet periods of an evaluation. It seems that they go through periods when they can't get anything done, and just can't get started. They, different from those who can't handle freedom, will often feel quite guilty, some even going so far as to suggest a reduction in paid time during this period. They are in need of support and assurances that the moratorium will pass and that when "the work crunch" comes they will more than make up for lost time.

110

Move Into an Alien Environment

I will begin this section by paraphrasing William Goode who, in one of his books on occupations and professions, says something like the following:

> Men at work and forests appear peaceful but upon close examination one finds that in both [work and forests], struggle is both swift and deadly.

It would be nice to believe that evaluators and agency personnel could work together happily and productively with little or no conflict, but that seems rarely to be the case. And it isn't that lined up on one side are the "good guys" and on the other "the bad guys," or that one set of activities are reasonable and another unreasonable, or that which one group is doing is more important than that which the other is doing. In fact, "good guys" are on both sides, both sets of activities are reasonable, and both important. The problem is that agency personnel, whether knowing it or not, turn power over to evaluators when they contract for an evaluation. While it is unfortunate that this is rarely made explicit when the contract is made, and even more unfortunate that it is only barely understood when it is made explicit, this transfer of power is a powerful determinant of evaluation-service agency relationship. Let me give an example. If an agency decides to do an experiment, the administration will impose restraints on the discretion of administrators to transfer personnel, start new programs, reallocate equipment, adjust schedules, etc., etc., etc. It is immediately apparent what this does to the formal power structure of an organization. Just contemplate for a moment on what it does to the *informal* power structure. And, the evaluator becomes, at times, the "tattler" and depending upon circumstances, at other times, the "enforcer." (It should not be surprising that in the eleventh month of a year's experiment even the chief, or top administrator, will want to give in to his subordinates. Often then only the threat of loss of external funds can assure completion.)

This conflict is compounded by the fact that often evaluators have different norms, goals, and lifestyles than agency personnel (this is especially the case for evaluators who deal with police) and it is possible for mutual "culture shock" to develop. The evaluator is often not used to the 9:00 to 5:00 day of many agencies. As a student he/she found that the computer was less expensive and more accessible *after* 11:00 p.m. His/her work patterns were made more tuned to his/her own personal rhythms than those of an organization. Bureaucratic niceties seem irrelevant. Adjusting to political realities seems dishonest. And so it goes. Both evaluation group and service agency find the work and lifestyles of the other alien. And little can be done to change that. Both staffs can learn to respect and tolerate each other, but only if they understand that conflict is not to be avoided, but rather managed.

So far I have talked exclusively about the necessary characteristics of field staff members of an evaluation. I would like to talk briefly about key characteristics of project directors. (Just as in the previous discussion, I shall be talking about the ideal. I am certain that just as perfect field staff do not exist in nature, so neither do perfect project directors. The extent to which I, as an evaluator, approach the following characteristics is unclear. I will not burden you with my own assessment of how I rate in striving for the ideal.)

Although I think other characteristics are important, I will identify three key ones: organizational "smarts," familiarity with the service delivery system, and experience in completing a project. I will keep comments about these to an absolute minimum.

Organizational "Smarts"

To me, administration and intra-organizational work is, to a large extent, the effective use of power to get particular tasks done excellently and then distribute fairly the benefits which accrue from getting the job done. Lined up against the struggle to get work done excellently are the work patterns, procedures, and organizational rules of grantors, sponsoring agencies, review groups, evaluation agencies, etc. Think of many of those for a moment.

Planning periods are not allowed. Generally a program is funded and started and then the evaluators are called in. False starts are not allowed. If, as in Kansas City, a false start occurs, most often the response is to "make do" rather than start over. (Read "make do" as "waste all the money, not just part of it.")

Failures are not allowed to be published. Rather than publish a failure so that other people

can learn, the tendency is to squelch a failure (so that other people can also fail).*

Decisions are not allowed. Often the administrator asks the question, "What does the rule book, organizational manual, etc., say?" The obvious conclusion is if the rule book says it can't be done then it can't be done. (What marvelous freedom for the administrator! All the prerequisites and none of the decision making.)

(Let me apologize to those of you who consider me outrageously irreverent in my attitude towards organizational rules and procedures. I have become convinced that the purpose of most rules is twofold:

1. They are to protect against "rip-offs"—although I suspect that more often than not, they serve to stop the very minor expense account "rip-offs" rather than the really gross ones.
2. They protect administrators from having to make decisions.

But let me add, it would be an over-simplification to say that procedures and work patterns ought to be removed. They ought not to be. They serve an important function. When properly administered they can protect agencies, grantors, etc., from gross rip-offs and absolute incompetence. Unfortunately, the rules, etc., do little to encourage excellence and can interfere with such achievement. The key is that an effective administrator has to learn how to wend his way through such rules, using them, if possible, to his advantage in getting the tasks done. There are various strategies to do this. I have known and seen "creative bureaucrats" who work 9:00 to 5:00 hours, take breaks and lunch at precise times, and who, because they know the rules and play the rules better than anyone else in the organization, use those rules to get jobs done. They are beautiful to watch because they have really mastered the skills of bureaucracy and remember that, ideally, the function of rules is to get a job done. [I have also seen accountants who understand that money is to spend to get a job done. Not spending money is no merit. It can be irresponsible *not* to spend money.] There are strategies other than being a "creative bureaucrat," but the skillful administrator learns how to use rules to *his/her* benefit. These skills are developed, honed, tested, in the *world*. They are not taught in universities and rarely talked about in bureaucracies. Learning them is accompanied by the acquisition of bruises, welts, scars, burns and *age*. Age alone doesn't do it, but it is only through the attainment of experiences to be reflected upon that these skills can be acquired. There are mentors and tutors to be had, but they rarely formally teach. *Most often they put you through it.* At early stages of your career you know only after you've been through a particular lesson and you sit bruised and smarting that you have been taught. Later, you know as it happens, and while you may not particularly enjoy it at that time, you can admire the skill with which it is accomplished. [But if you have concentrated during your early lessons, there really aren't all the accompanying pains, just generally the reminder that when doing complex work it is necessary always to be very alert.])

The coupling of energetic, bright, relatively undisciplined young researchers with a seasoned organizational veteran who can provide a certain amount of structure (or the appearance of structure) seems to me a likely guarantee of a reasonable success in completing an evaluation.

Familiarity with the Field of Service Delivery

While I am not sure the following assertion will be absolutely clear, I nevertheless want to begin with it. I am *not* interested in evaluating particular programs. I am interested, and I think my clients are best served, *if I evaluate methods and strategies, not programs.* Let me explain that. The important principle here is generalizability. A program is only of general interest when it exemplifies methods, skills and strategies which are relevant to a wide variety of settings. Programs may or may not be that generalizable. If a program is so dependent upon local circumstances that it cannot be exported to other settings, I, as an evaluator, am simply not interested in it. It may be that it is of legitimate interest to the agency program officer. But I am interested in developing the knowledge base about the effectiveness of methods and strategies which are transferable in a broad field of service delivery. In order to see the broad application of a project, an evaluation director must *know* that service delivery system, must be aware of the intellectual traditions that have given rise to the present knowledge and skill base of that profession. And, it seems to me, she/he must be able to help the client context her/his program in those traditions. If the evaluator can't do that, outcomes are meaningless.

I did not include this in the characteristics of evaluation staff. If they would have such knowledge of the field when they started, that clearly would be desirable. But it is not essential that the evaluation director makes certain that staff acquire it during their work. Staff will, if highly motivated (one clue to the curiosity, skill and interest of an evaluation group is the extent to which they quickly start immersing themselves in the literature to acquire familiarity), acquire familiarity with service theory in relatively brief periods of time. (Methodological sophistication cannot. That has to be learned by doing as well as studying.) But since the project director is the person who will be set-

* This is really a very complex issue and one that can only be referred to here. The publication of failures is dangerous to agency administrators because it simply provides another weapon to those who are always lurking in the wings, waiting to exploit any mistakes made by competent people who make mistakes and are willing to admit them. As a result the publication of mistakes has to be carefully orchestrated.

ting the general directions of the evaluation group and providing the overall guidance, it is essential that he/she know the substance and theories of the field.

Experience in Completing a Project

Evaluations don't complete themselves. A staff can be skilled in data collection, analysis, theory building and grantsmanship and still not be able to complete an evaluation. The best of people can block in completing an evaluation. It's almost a stage in research or evaluation. The person who has been through completing a project knows the project can be completed. The fact that at least one person knows it can be completed is critical. Outlines circulated widely to colleagues and consultants can help disperse the feeling of hopelessness which develops when people sit down to write after five years of work and $600,000 of funds. And, if they have kept their records, exploited the resident obsessive compulsive, and if they can narrowly concentrate on the questions the program addresses rather than the "oh by the ways," the first rough draft is half written by the time they sit down to write. (In other words, if the project has been well run, the writing of the final report began with the development of the original grant. Report writing implements include scissors, scotch tape, xerox machines; as well as pencil and paper.)

These then are the characteristics that I find essential in good evaluators, both staff and director. No doubt there are other characteristics which should be addressed here, but, at least for me, the mentioned ones are most critical. *This is most certainly true.*

Other miscellaneous Verities:

Verity #3. **In order to understand one (police officer, physician, nurse, social worker) you must *not* be one (the other side of—"In order to understand one, you must be one")—or—*cooptation.***

Much police, social and medical work is perceived of, and often is, exciting and important. For young persons who have hardly seen the outside of a university, such real world work will be attractive and interesting. For many it will be a welcome relief from the years of thinking and reading rather than doing. Their high degree of interest in such activities makes them especially vulnerable to cooptation.

My own experiences have led me to the following points of view regarding cooptation.

1. It is to be expected. It is a stage that all researchers must go through if they are properly sensitive to their subjects.
2. Cooptation is a trade-off. Whether agencies and evaluators do it consciously or unconsciously, both try to seduce the other to their respective points of view. In so doing, both allow an unusual amount of access to the "secrets" of their organizations. When remission from cooptation occurs, the researcher (or professional) is generally much wiser about the other organization and him/herself.
3. Although there are counter-strategies, i.e., supervision, and creation of a staff culture, most often remission is spontaneous and occurs when a terribly biased initial report is reread with horror and shock several months later. (Here, good supervision points out the universality of the ailment, is supportive, and recognizes it as an important learning opportunity.)
4. There is no subsequent immunity to it. It happens over and over, even to crotchity old project directors.
5. If remission does not occur, more likely than not it is terminal and career counseling is in order. Unreconstructed co-optees are a disaster to evaluations. They are devisive, secretive, and generally have all the zeal of religious converts. Truth is theirs alone.
6. Symptoms include: (for police evaluations—people doing evaluations in other agencies will have to fill in their own specifics)
   a. Wanting to carry a gun.
   b. Feeling that nobody really understands the police as well as you do.
   c. Becoming a police "buff."
   d. Overemphasizing confidentiality. (When cooptation has occurred, the principle of confidentiality includes and more often than not is specifically targeted at the project director. The researcher feels that he must "protect the poor police department and police officer" from the rapacious project director.)
   e. Developing the police "swagger."
   f. Using police jargon.
   g. Wanting to get involved in the action, i.e., help with arrests, etc.
   h. Ignoring findings or "twisting the text to meet the message."

And finally, I would argue that the staff member who is never cooptable simply is too disinterested or too far removed from the issues. Cooptation is like sex and love relationships. You might not want it all the time, but without it there's boredom and disinterest. *This is most certainly true.*

Verity #4. **The only truly unforgivable sin is covering mistakes a second time—or—*mistakes at work.***

Mistakes are common for people at work. My own feeling is that I make a minor mistake a day, a middle range mistake every week, and a truly major goof-up once a month. Such is the nature of work. But mistakes are not to be confused with in-

competence. People have rights to mistakes, but not to incompetence. And the nature of the world of work is such that, given proper colleagueship, supervision, and direction, most mistakes can be handled and compensated for—most often by extra work. (That is to be expected.) And while it might sound Pollyannaish, I really believe that mistakes and the handling of mistakes provide some of the most critical opportunities for learning and growth to capable reflective people.

Further, it is to be expected that some persons who make mistakes will try to cover them up (not by redoing the task but by hiding what they know or lying). As a result, a project director has to be careful to remain familiar enough with what is going on to be able to spot the covering of a mistake, especially a major one. When "covering" does occur dramatic action is necessary. All must be made to know that that is the one unforgivable sin and, if "covering" ever occurs again, that's it. Termination, firing, is the only alternative.

But, for the most part, mistakes simply have to be lived with as a fact of life. Often one can only shrug off the minor mistakes knowing that it would be more of a mistake to try to undo it than just to forget it. The middle range mistakes often have to be made up for by extra work (not that anyone tells you you have to, it's simply work that has to be corrected). Regarding the major mistakes, they not only require effort to undo (some may be so serious that they cannot be redone) but they also provide rich learning experiences in living with the consequences of life. Be clear, major mistakes generally do have consequences, but most often the consequences are not calamities if faced up to.

For me, my primary goal regarding my own mistakes is to discover them myself and report them. (This can be read as honesty or practical realism.) Such reporting does not free one from the consequences however. It simply is the development of trust in work relationships. I hope that my boss can trust me completely. That is—that he can trust that I will make my mistakes, but that he will never be surprised by them. I have found few mistakes that cannot be handled in civil ways. Covering a mistake, on the other hand, may mean that the opportunity to redo it is lost and potentially is disastrous to a project. (If I sound "preachy" at this point, it is because I feel quite strongly about this. Much of the work we do in evaluation is new and exploratory. If staff runs scared because they are fearful of making mistakes or taking appropriate risks, then the whole enterprise is lost. Evaluations are simply risky business. Bright competent people have the right to mistakes. Evaluations and evaluators can fail. If failures are seen as legitimate, then we can continue to develop our field, both through the successes and failures of ourselves and our colleagues. But failures, too, should be published so we don't have to go on and on making the same major mistakes in evaluations.) *This is most certainly true.*

Verity #5. "Identifying the laborer who is to be in the vineyard"—or—*selecting a subcontractor.*

Although I do not have a great deal of empirical evidence about this, I nevertheless am convinced that every evaluative organization has a genius of design working someplace in the inner sanctums of the organization. That person is not only a genius but often too has E.S.P., in that she/he seems to be uncannily aware of exactly the design the contractor has in mind. But the grantor will never meet this design genius and once she/he has completed the design, she/he will be irrelevant to the evaluation. The point I am making is that the key persons to assess in selecting evaluators are the people who will actually do the work. They will make or break the evaluation. Even the project director is not enough. You must see and make judgments about the key on-site evaluation staff member(s). *This is most certainly true.*

Verity #6. The truth shall make them free—or—*passing by the crotchity old evaluation director.*

And finally, if young researchers are bright and capable, and if an evaluation director has given them the opportunity to really use their magnificent selves and skills, and if he/she believes that knowledge and skills are really crescive, the evaluation director will see young evaluators fly slightly higher and slightly faster than the crotchity old evaluation director. And that's what it's all about and *is most certainly true.*

Conclusion

Those of you familiar with hermeneutical principles will recognize that I have used the classic three point Lutheran sermon style: Introduction, three points in the body with the central part being both the longest and most important, and the third part a miscellaneous section where things are put that don't fit into the outline. The conclusion is generally an exhortation. I have presented my verities. I shall spare you further exhortation. *And that is most certainly true.*

One final point. My evaluation colleagues, the Kansas City Police Department and I have completed an experiment which has been considered to be fairly well done. We were very, very lucky. We worked very, very hard. Most of the things I am telling you are in hindsight. I may be wrong. I think I am right. *That is most certainly true.* Selah. Amen.

Additional comments on putting together a good evaluation research team.

Lee Sechrest

The skills involved in carrying out good program evaluations are special and not widely available. There are sufficient special characteristics of program evaluation research to make it unlikely that researchers without specific experiments and/or training for evaluation will be able to resolve all the problems that are sure to arise. Therefore, an administrator wanting to become involved in program evaluation research will not maximize chances of successful completion of the evaluation by relying on the usual sources of research expertise in his community, e.g., a local university faculty. Unfortunately, many university faculty members have no notion that their capabilities may be in any way limited.

In fact, most administrators will need some help in locating and recruiting evaluation researchers. There are several sources for such help. First, the potential funding agency for the research will often know a good bit about the local research community and will be able to make recommendations based on their experience of researchers who have the needed expertise and interest. A second source of information often available is the directors of other similar evaluation research projects. If an administrator knows of evaluations which he or she considers to have been well-done, a good move would be to contact the evaluators of those projects for advice. Even though the evaluators are at a considerable distance, evaluation researchers will often know the resources available in the community. Finally, the administrator may inquire locally to determine whether there are evaluators with experience of the type needed. The administrator should not be reticient about asking to examine credentials and samples of previous evaluation reports. If necessary outside help, e.g., from funding agencies, should be sought in assessing the credentials and previous work samples. No competent and honest evaluator will balk at having his or her work examined carefully.

A good evaluation research team begins with a highly competent evaluation researcher. That person will then, ordinarily, be able to put together the staff to the evaluation if it is funded. In the meantime that researcher should be quite willing to participate in planning the evaluation study and in preparation of the proposal to be sent to the funding agency. The greater the input from the potential research director, the stronger the proposal is likely to be and the greater the chances of the ultimate success of the evaluation.

Evaluation of Experiments in Policing: What are we Learning?

Joseph H. Lewis
Director of Evaluation
Police Foundation
Washington, D.C.

Over the past several years the Police Foundation has been fostering, supporting, monitoring, and publishing results of a variety of research on the delivery of police services. During that time the Police Foundation has accumulated a valuable fund of information about the problems in doing police work and in getting it paid attention to in the police community. While police work cannot be equated with the delivery of emergency medical services, it is believed that there are enough similarities between the two fields to make at least some of the lessons learned from police work transferable.

It has been a long time since I have done any work, but I have had the opportunity to learn from the labors of others. The last five years have been especially interesting. During that time the Police Foundation, in collaboration with a number of police agencies across the country, has initiated fifteen substantial pieces of evaluation research in the field of urban policing. Ten experiments are finished, three are in various stages of evaluation report completion, and two are still running.

Some experiments have been done by Police Foundation evaluation staff with support in some instances by contract research institutions, many by research institutions under direct contract to the Foundation. These numbers do not sound impressive compared to, say, the national debt, but they do, in fact, constitute a respectable fraction of the evaluation research in regard to policing that can be termed consciously formal in the sense that it is intended to conform, as far as nature will allow, to the rigorous standards of science. Since these are a class of social experiments we are talking about, it will come as no surprise to you that sometimes the correspondence with scientific standards of rigor has not been as close as one could wish. But all of our work has been conducted, reviewed and reported by those standards.

Much that the Foundation does is of a different nature, related to removal of barriers to improvement in personnel and other important aspects of administration or to more direct efforts at reform through information exchange and the like, but all of the activities under direct discussion here were initiated with the firm intention of formal experimentation. Each initiation has been the product of a negotiation between the Foundation and a police agency. Each negotiation began with exploration by a Foundation program officer with police administrators to search out possible issues of common interest which lie within the strategic purposes of the Foundation, in policing situations that appear to lend themselves to productive research.

When an acceptable issue to test is found in a climate of circumstances that appears to favor formal experimentation, the program officer works with the police agency to help the agency to produce a proposal sufficiently concrete to enable our Board of Directors to assess the intrinsic worth of the idea, in terms of generating nationally, as well as locally, usable knowledge of substantial importance to improving policing, and to consider the cost to develop a program plan for the experiment and an evaluation design to go with it. This preliminary proposal will have had, at the very least, input and advice from me with respect not only to evaluation design and planning needs, but also about bringing the statement of program purpose and process toward measurable, concrete terms. Often, even at these very preliminary stages of program development there will have been more extensive evaluation staff collaboration in specifying what kind of experiment it will be attempted to design.

When the Board approves the planning grant and a sum for evaluation design, the police agency adds officer and other capacities—including civilian professional specialists as needed—to the planning team which will develop the full experimental design and program of action. Evaluation capacity is mobilized to work in close conjunction with the planning team to produce the evaluation design and work plan so that the experiment and evaluation are parts of a single, coherent entity aimed at producing the defined knowledge specified.

Initial estimates of the experimental design task, of the capabilities of the police or evaluation groups to perform, or both, may have been mistaken. If the design and planning process goes well

but needs more time or other additional resources, extensions to as long as one year, on one or two occasions even longer, may be funded. If it should become clear that a feasible design for formal experimentation and evaluation is not going to emerge, no experiment will be funded. Should another kind of research than an experiment still seem promising, a proposal for it, prepared through the full cooperation of the police and the researchers, would be submitted to the Board for consideration.

A grant to a police agency to conduct an experiment or other form of research requires the agency to commit itself to facilitate collection, and in some cases to provide, baseline and other data pertinent to maintenance of the experiment and conduct of the evaluation. It must also commit itself to maintenance of experimental conditions for the planned duration of the experiment, barring catastrophe. Foundation program officers monitor and work with the project management staffs of the police agencies in which they have experiments or other programs in progress to make sure that the agencies have the capacities needed to maintain controlled experiments and are doing so. Should that not be the case, every attempt would be made to assist the agency to do so. If circumstances did not allow for full success but the agency remained committed to the attempt, adjustment of objectives might be made if substantial gains in knowledge could still be expected. Otherwise funding would be subject to termination.

These, no doubt, simple appearing paragraphs compress a great deal of information about what we have learned about doing evaluation research in policing. It is the model we believe to be most useful in our business. We have come close, much closer perhaps than most, to operating as I have described. Even when we do, there are serious problems to deal with.

Development and conduct of experimentation and evaluative research in these fifteen instances has provided rich experience in identifying some of them. Several of your speakers are participating in this conference because Professor Sechrest believes that some of our learnings from them may be transferable to research in the field of emergency medical services. Our practitioners and researchers in that field can assess which ones may be applicable and to what degree that may be so. I shall not myself attempt to draw many parallels. There are probably many reasons why I should not, but one seems sufficient: I don't know enough about emergency medical services (EMS).

Let us begin to unravel some of these generalities. Note first that all of the foregoing has been stated in terms of the interests of a funding agency, one dedicated by the terms of its charter and commitment from the Ford Foundation in late 1970, to improvement of policing in the United States.

There are a number of reasons for this. An obvious one is that that is the perspective natural to my present business. Another, however, of more direct interest for this discussion, is that the funding experience can be a sort of integrative mechanism for learning. When we take note over time, for example, of what the most useful items are that our funds provide with respect to initiating or to sustaining an experiment or an evaluation, or to keeping them in adequate relation one to the other, we begin to understand which of them seems special to one circumstance and which are recurrent and more general in application. It is the fact of being a funding nexus that lets us learn the same thing across a variety of projects about the importance of what our program or our evaluation people do. Once we have understood those observations, the findings that seem to be most general can be used by any agency that wants to test, in a formal sense, the usefulness of what it already does or innovations that might improve the agency's effectiveness.

Finally, this perspective is suggestive of another important point. When the Foundation was first chartered, it was expected that a flood of good ideas about things to try, expressed in terms of well thought out and specified proposals, would pour in from police agencies across the country. A flood did pour in at first, but in general, they were requests that the Foundation fund conventional training programs, or a new headquarters, or a management survey or the like. Those that referred to a desire to try a new idea often showed an unawareness of what other agencies were doing or were not well thought out in terms of specified objectives, concrete steps to achieve them or measures of success. In short, it quickly became clear, even to those of us who did not already know it, that the Foundation was never likely to be able simply to hand a check to a police department and stand back to wait for the inevitable good results.

The problem for the police is that they are fragmented into some 17,000 forces, each an island unto itself. They can be islands in two senses important to this discussion. They have tended often, as you probably know, to feel defensively isolated from the communities they serve. In cities where our surveys have shown, as they invariably do, that citizens have a high regard for the police and are supportive, the police tend to underrate that regard and support. There is an aura of secrecy about what the police do and how they go about it.

But, for our purposes, almost more important is the fact that police agencies are, generally, insular with respect to each other. Almost all of our nearly half million police serve their whole careers in the agency they first join. Lateral movement except at the highest levels is almost non-existent and is rare even at the level of chief. Communication among them about the substance and methods of

their work is generally poor. In the spring of 1974, the Foundation convened a conference of the chiefs of patrol of the forces in the 35 largest cities in the country. That is the first time they had ever met.

These factors seem to have had consequences of the following kind. It is rare for police administrators to be formally trained in management, as city managers must be, or in business management. It is rare for police agencies to employ the many professional or technical skills from "outside," as many other forms of enterprise that deal with organizational management and human service issues find it natural to do. Management practices common to many other forms of enterprise are slow to be adopted in policing. State-of-the-art knowledge or breadth of experience with problems and practices across differing jurisdictions is hard to come by in such a setting.

This is why the money the Police Foundation provides in planning grants goes largely for two things: "outside" consultants and travel.

Over the last few years we have helped several police agencies learn how to use psychologists, sociologists, program analysts, data technicians, personnel specialists, organizational development specialists, and others with talents and specialties from outside the world of policing. It has been necessary to do so to help police administrators formulate in concrete terms the ideas they want to join with us in testing, to help them learn what else is known that is related to it, to help them select the most promising ways by which to test their ideas, and how to make those tests acceptable, with meaning to patrol or other officers, as well as to the citizens, who are affected by the test or who may be by the results.

Travel budgets for other than the chief are small or non-existent in many departments. Even the chief may be restricted to one or two trips per year. Travel is often the first item to be cut in tightened city budgets. A cutter simply has to say "boondogle," and wield the axe.

The Foundation has sponsored travel, by officers at all levels, to other cities that have dealt in some way with an issue area they wish to explore that will help them in their planning.

Some have said that providing these two kinds of aid to police agencies, helping them to open up to a broader world, both of policing and of the still wider one beyond, may be among the most useful things the Foundation does. I would not deny that possibility. It is, at any rate, clear that we could not design and plan good research with our police partners without them.

Does any part of this sound familiar to you as EMS practitioners and researchers?

Let us move on now from what we have learned about what it takes to help a willing police agency design and plan good research to what we have learned about what it takes to execute a good research design to produce credible answers about what works or what does not. To lay the ground work, consider what we need to deal with.

Evaluation of the consequences of experimentation requires, ideally, commonly accepted, well defined measures of input and output. Measuring the performance of police requires agreement about the objectives of policing, what the police are supposed to deal with, how they are supposed to behave, and what they are supposed to accomplish, all in measurable terms and based upon data that it is feasible to get. It is common knowledge that measurement of public sector activities is generally far more difficult than for business where dollar gains and losses are comparatively easy yardsticks to apply. Policing provides an excellent illustration of the complexities of measurement in the public sector.

Let us trace that idea for a moment. One origin of the problem is that there generally is not one public which decides and transmits through city management what it wants the police to do; there are several and they are often in sharp disagreement. Field interrogation, stopping and questioning citizens, can be proper order maintenance to some middle class blacks or whites and, at the same time, harassment to youngsters with long hair or bushy afros. Some want and need emergency helping services, from transportation to medical service, to counseling about domestic trouble, to solving neighborhood disputes, to dealing with an insane relative or friend. Others in the same city would turn to their doctor, their marriage counselor, their lawyer, or their psychiatrist for these sorts of service, believing firmly that the police should "stick to crime" or "solve the traffic problem" and not be diverted by these, as they would term them, extraneous, unproductive demands on their time. And so it goes.

For any particular remedy the police might apply, there will be disagreement about its use. Is an arrest the best solution to a problem? People differ. It is almost automatic for many in and out of policing to think of good policing as aggressive policing and to think of high arrest rates as indicators of good, aggressive policing. But for several years, many have thought not for some kinds of behavior the police most often deal with and have tried to divert young offenders, or drunks, or others away from the law enforcement system, or they have wanted to teach police to counsel police in domestic disputes, partly so as to avoid arrests whenever possible. Some believe the police should be cool and impersonal, others, warm, friendly, interested.

What this means for research and test is that no single measure of performance or outcome will suffice. As many aspects as possible must be measured and the results laid out so that any police or public reader may apply his own relative weights or values to them.

Another perspective that helps to understand why evaluation of experiments or assessments of police performance or of effectiveness are complex and difficult stems from recognition that little is firmly known about cause and effect relationships in dealing with crime, little theory exists that explains how or why what the police do ought to affect crime. Only a tiny beginning has been made. Two examples will help to make the point. It has been assumed as a rule by many, in and out of policing, that one-third to one-half of the time of police officers assigned to street duty must be spent routinely patrolling the streets to prevent crime, insure citizen satisfaction with the police and reduce their fear of crime. Our experiment in partnership with the Kansas City Police Department [1] suggested that quite wide variations in routine preventive patrol, keeping everything else constant, had no effect on crime, satisfaction, or fear that we could find. Another Kansas City experiment [2] that the Law Enforcement Assistance Administration is funding is beginning to suggest that, in many instances of even serious crime like street robbery, citizens wait so long before they call the police that it does not matter whether the police hurry or not as far as opportunities for on-the-spot arrests are concerned. And yet both police and public have always felt sure that short response times were good for that. In fact, short response times are often used, by themselves, as indications of a good police force. And police managers coach their publics to expect short response times to all kinds of calls and they spend substantial resources on radios and cars, manpower and computers to make them short, an expensive proposition.

What this says is that there is not yet much validated, codified knowledge and that much of what we think we "know" is not true. Clearly, then, in the field of policing it is important to test the conventional wisdom as well as to try out new ideas. We must expect our lack of knowledge to complicate our research designs and to increase the risk of failure for unexpected reasons.

The effects we are looking for are often subtle or modest in size. The measurement tools so far developed are not always very sharp. Many believe that, to some unknown degree, much criminal behavior stems from economic and social conditions. Young people are being arrested for a large and growing amount of it, up t half in many places. The police cannot keep people from being young, or poor, or male, or black. What police can do can affect some kinds of criminal behavior some of the time in some places. When we try to use the amount of crime reported to the police, and that the police include in their records, to determine whether crime is changing, we run the risk that any changes we may see may be caused by differences in what people choose to report to the police. They may also be caused by changes in the way the police treat the reports coming in. These problems can be guarded against for certain kinds of crime by conducting victimization surveys of citizens. Data from such surveys do not have police bias in them but surveys have some problems of their own. What looking for modest effects with imperfect measurement instruments demands is measurement of any given effect from as many perspectives as possible. Such multiple perspectives when applied to a sizable number of outcome measures can give confidence about what did or did not happen even though, taken singly, most measures would be too weak to do so.

But it is not impossible to bypass all of these complications by noting that, since the business of the police is to provide service to the public, direct measures of citizen satisfaction with police service would be the ultimate indicator of success or failure? Unfortunately this is not now a real possibility. If the lack of hard knowledge and the other complications we have mentioned are linked back to the earlier point about insularity of police with respect to their public and the secrecy that surrounds what they do and how they do it, the result is that citizens have little or no basis for knowing what it is reasonable to expect their police to accomplish or how to judge whether how they go about it is productive or wasteful. This denies evaluators the straightforward use of indicators of citizen satisfaction as a measure of adequacy of police performance or effectiveness.

What have we learned about conducting research, experimentation and evaluation in such an environment in partnership with police agencies? Let us go back to the condensed summation with which we began to see what those simple looking statements mean in practice.

We said that a police agency that wants to test an idea must commit itself to facilitate collection, and in some cases to provide, baseline and other data pertinent to maintenance of the experiment and conduct of the evaluation. The importance of baseline data, that is, data that shows what conditions are before a contemplated change is begun, is pretty clear. Without it, it would not be possible to make serious before and after comparisons to show whether any change took place when a new technique or other change was tried. But what many administrators whose experience has been concentrated on operations, making things happen, are not prepared for is that collecting such data can be a massive, time-consuming affair. Commonly, it has been their experience that it is difficult to gear up their organization to generate support for change or innovation, or to challenge accepted wisdom. We will come back to this point

[1] George L. Kelling, Tony Pate, Duane Dieckman and Charles E. Brown, *The Kansas City Prevention Patrol Experiment* (Summary Report, 1974; Technical Report, 1975), Police Foundation.

[2] Deborah K. Bertram and Alexander Vargo, "Response Time Analysis Study: Preliminary Findings on Robbery in Kansas City," *The Police Chief*, May 1976.

in a moment. Once that enthusiasm has been generated, it is natural to want to act before it dissipates. What has to be done in practice is to incorporate that baseline data collection process as an integral part of the agency's preparations for the experiment. It is easier to do so if the issue to be addressed by each test is as concrete as possible. The measurement complications and lack of theoretical knowledge of policing to which we have previously called attention also suggest this course.

The process of bringing an organization to the pitch of enthusiasm often generated to facilitate launching and support for maintaining an experiment or other kind of innovation in policing can result in a state of overpromise leading to subsequent disillusion. It is something like the politics of congressional legislation, so much has to be promised to secure passage that any action bill is almost automatically doomed to be seen as a failure when it is implemented. We noted earlier that most of the effects the police can produce by changing what they do are expected to be modest in size. Overpromising is easy, disillusionment—both of officers and of the public—is frequent and makes further change more difficult. The shrewdest chiefs have learned to focus on the trying of better ideas or the testing of old ones to make improvement rather than on expectations of eliminating crime or citizen fear by any single thing, however major, their departments, by themselves, can do. This is a hard-learned but valuable lesson for other managers of service systems.

We also said earlier in our initial summation that a cooperating police agency commits itself to maintenance of experimental conditions for the planned duration of the experiment, barring catastrophe. Let us deal with catastrophe a little later. Experiments do not maintain themselves. By definition, they constitute the maintenance of strange conditions. Organizations have enormous capacities for absorbing attempted change so that when one looks again, all is as it was before. There are many reasons for this: Practitioners may believe that the way they normally do their work is best; they may feel that a change to be tested risks the safety of their beat; individuals may fear a loss of relative power or prestige, or even pay. Collectively the effect is similar to inertia, an organization tends to keep on doing whatever it has been doing in the same way it always has unless an inside or outside force is brought to bear to change it.

To be serious about research that requires experimental conditions to be set up means that the police administration needs to decide in advance how it will know that those conditions are in being and to set up explicit means—data or indicators to watch and people to do it—for continuously or periodically monitoring whether they are. Such a monitoring capacity must be able to feed information to the boss as to what is off the track and what change will restore it. It is then up to the boss to take the necessary action to do so. If some police activity is to be stopped in defined areas, is it stopped? Does it remain so? If an activity, or the number of officers is to be increased at certain times or in certain areas, is that happening? If two kinds of officers, say male and female, are to be assigned to tasks equally, in this case without regard to sex of officers, is that being done, or are men subtly protecting women?

In practically every case, the cooperating police agency has required the continued internal assistance of some of the same kinds of consultants that were provided to help with the initial design, and planning of the research. To these have been added police management and operational talent which together form a program management group to run the research program on behalf of the agency.

Often, and what the Foundation especially likes to see, the city government, at the recommendation of the police administration has created the necessary budgeted positions to institutionalize the civilian additions to the police agency's capacity to plan and manage research after the first year or so of Foundation support. Such bodies often assume wider planning, and grants and research management capacities that carry the agency's ability to innovate and test what it does well beyond the initial levels the Foundation has sponsored. The Kansas City response time study was designed, funded and conducted, including the presently ongoing analysis of results, through the efforts of the research capacity originally established in the course of Police Foundation experimentation in that department.

We had said that Foundation program officers monitor and work with project management staffs to make sure that the agencies have the capacities needed to maintain controlled experiments and are doing so. The energy and attention of our program officers have often been as important as our funding in securing the successful completion of research. When the indicators show that some condition is not being maintained as agreed, it may be that a shift of existing program resources will help to get it back on track. A staff visit to another department where a similar problem has been solved may help the agency's project management more than additional computer time that may be budgeted. Or a computer specialist may be able to solve a programming problem to help get better data for controlling the experiment. Flexibility in shifting experimental program resources has often helped to make the most of research opportunities.

The police agency's own monitoring process is designed during the early planning phase that we have talked about when the experimental and evaluation designs are being worked out together. The evaluation team works with the agency's proj-

ect management staff and helps to specify what indicators will show whether the experiment is on track and assists in designing the data collection scheme that will produce those indicators. Once the experiment is running, the evaluators monitor the quality of the indicators and help the agency to improve the quality where it is not adequate for the purpose. In every case so far that has been necessary. One reason is that data adequate for every day familiar operations are often not sufficient for doing research or trying out new ways to do things; the level of detail may be too low or not all the kinds of data needed may be routinely collected. Another is that many police agencies are in some state of transition in their use of computers. This means that, even though the computer is producing data about an operation, the operation may still be being managed and run by the pre-existing method of control. In such cases, errors in the computer data may not be noticed. In any case, they do not matter. When, for example, adherence to dispatch discipline in a team policing experiment forces use of computerized dispatch data, errors in the data suddenly make a difference. Before that, no one knew that there were any.

The four-way feedback between police agency program management and evaluators in the field, between police and Foundation program officers, between evaluators in the field and evaluation management and, finally, between Foundation program and evaluation management, has been responsible, at its best, for getting the most out of a research opportunity to help a police agency gain knowledge about a question important to its own purposes, as well as to policing nationally. When communications in one or more of the links has been incomplete or slow, results have tended to be less satisfactory. This may happen because the capacity or behavior of the police agency or evaluation staff could not be adjusted rapidly enough.

When circumstances beyond control prevent realization of initial expectations for an experiment, it is sometimes true that less ambitious but still valuable research objectives can be reached if the facts are learned soon enough that police agency and Foundation management, both program and evaluation, can agree on the changed research specification. If events preclude that, it is still essential that these feedback loops, especially from evaluation staff, operate so as to make clear to all concerned how a given state of affairs differs from what was planned. For example, it can happen, as it can in most public or private bureaucracies, that a prime source of inertia or resistance to change is middle management. A decentralization plan, perhaps such as neighborhood team policing, when implemented, will shift operational decision making authority downward away from middle management. If other aspects of the change in organization and operations do not compensate for that in ways perceived as adequate by middle management, members of that group may well resist the maintenance of the new arrangement so that, in a few months or a year or two, authority they deem important will become re-centralized and the planned change will really not exist except, perhaps, for superficial appearances. Should such a state of affairs be detected, it would be important for a chief to know as soon as possible so that he could decide whether he has the political power, internal and external to his agency, to deal with the situation. (We will come back to this point again a little later.) If circumstances change, it is important for all concerned to know that the evaluation report will say that.

By now we have seen that, in all cases, operating agencies have added new capacities to themselves to enable them to plan and conduct serious research. The sorts of capability adequate for operating as usual are not adequate for an agency that really wants to advance its knowledge of, and to improve, its own art and practice. The sorts of additional talent that are needed do not ordinarily grow in police agencies so they must be brought in from outside where they do, from universities and research groups, from technical and professional schools, from other backgrounds and experiences. When this has happened, not only has the agency been able to conduct research and tests that it wanted to do, but also, it has been able to improve its knowledge and control, for management and operational use, of its data and information systems; it can analyze its own internal operations; it has been able to plan, secure funding for, and execute additional research and test or other improvement projects on its own. Most importantly, the viewpoint of the agency can change to one of open questioning of what it and other agencies do and how they do it, making learning from experience a continuous, explicit process, and innovation and change based upon such learning, natural. This is a sharply different atmosphere from the isolated, defensive, rigid climate which has pervaded agencies that have not moved.

Adding such capacities, even only one or two people bringing new kinds of talent and training not "slotted" in the organization, costs money. Sometimes part of the operating force or of management that is to participate in an experiment or other research need to be specially trained. That costs money, sometimes at overtime pay rates for large numbers of officers, plus the cost of instruction. Sometimes additional or special equipment is needed (although the Police Foundation has tried to keep its contribution to equipment at a minimum), and that costs money. City or county councils do not, even in relatively good times, readily make money available for research and experimentation; they prefer to fund only traditional or tested items. If it were not for that, police or other agencies could go ahead and add whatever abilities are needed and do their own research and

testing of what they do or of new ideas. As it is, with rare exceptions, outside funding sources must always pay the bills for initiating test and innovation. And another need for outside funding is to make evaluation credible.

It may seem strange that we have come this far in discussing evaluative research in policing with only cursory mention of evaluation. We have said that evaluation and program designs must be developed and planned together as parts of a coherent whole; that evaluation staffs help police agency project management staffs to design and test internal project monitoring and evaluation plans and data systems for them; that evaluation staffs monitor these monitoring systems and independently assess the state of maintenance of experimental conditions. We have said that evaluators provide crucially important feedback about that to the agency and to the funding source, to both program and evaluation managements. But that is all.

One reason we have not said more is that other speakers at this Conference have already done so. But the most important reason is that we are dealing with first things first. An agency chief and administration that really wants to test an idea, is fully committed to maintaining agreed upon experimental conditions for the duration of the test, has the capacity to design and plan a good experiment and the ability to monitor and to take whatever action is required to maintain it, can make the evaluation task, inherently difficult at best, worth trying. If the agency chief and his administrators, either through lack of interest or impatience, lack of understanding of the commitment they have made and what it may require them to do, or for any other reason, do not maintain the experimental conditions, the planned evaluation is impossible and no amount of evaluation talent can make it otherwise. So we have concentrated here on what service practitioners need to do to make experimentation and evaluation feasible.

Given that the conditions for research and experimentation leading to opportunities for good evaluative research have been established in an agency, why should it not go ahead and do its own evaluations? For many purposes it should. This will be particularly true for tracing of internal operating processes and attempts to change them and for some experiments which can be evaluated at relatively low cost. An ability to do so will not only enhance the ability of such an agency to do its own work better but will make it a much smarter customer for outside research it may wish to contract for—a point of no small importance when one is aware of how vulnerable most agencies are to the purveyors of outside "expertise" and how little unsophisticated agencies benefit from such services.

But factors work against the agency doing its own evaluations in many important circumstances. One is that if the agency wishes to make a substantial contribution to better understanding of a police issue that has national importance, it is essential that the evaluation of results of an experiment done for that purpose be, and, most importantly, be seen to be, disinterested. A separately funded, independently managed evaluation staff to measure impact of the conventional wisdom or new technique or operation being tested is essential to credibility, though even that does not necessarily assure it. That is why, in all experiments sponsored by the Foundation, the evaluation is funded by our Board in a budget entirely separate and distinct from the budget for the program to be evaluated; the evaluation capacity, whether internal to the Foundation or contracted for, is managed and directed entirely separate from program management, and both designs and draft evaluation reports are extensively reviewed by an outside Evaluation Advisory Group, members of which have no vested interest in the success or failure of a program or of a police agency. A more complete separation would occur if the Foundation sponsored the evaluation of an experimental program funded by others. This has happened but is rare, partly because so few experimental programs well enough controlled to be worth formal evaluation are being funded or are occurring naturally, partly because others who fund programs, not unnaturally, want to reap the potential benefits which may come with publishing reports of good outcomes. Since experience has taught us the, literally, crucial importance of program monitoring and control of experimental conditions, the separation of program and evaluation management but still within the Foundation rubric has seemed to us so far a most useful compromise between assurance of as high quality research as the situation may allow and the high external credibility of results.

The other reason why evaluation of experimental impact must most often be external is cost. It is not unusual that baseline data that must be collected even before it can be known that the experiment will run successfully, or even for sure that it will start, can easily cost $100,000. A completed evaluation of a major experiment, such as the Kansas City Preventive Patrol Experiment, conducted by Dr. George Kelling and the Police Foundation Kansas City Evaluation Staff, with technical support from Midwest Research Institute, may cost $650,000 to $700,000. The five-year, from start to design to publication of report, Urban Institute evaluation of the Cincinnati neighborhood team policing project known as ComSec will have cost well over $1 million when it is completed, this despite the effective efforts of Alfred Schwartz, who managed it, to keep the costs as low as possible. Such costs come about through the inherent difficulty of answering the questions

we are attempting to deal with, however simple they may sound, in the face of the complexities about measurement in policing to which we alluded earlier and with the rather blunt tools at our disposal. In order to say whether sex is a bona fide basis for exclusion of policewomen from patrol, it was "simply" necessary to determine whether some women could perform as well on patrol as acceptable male officers. Given the disagreements about what patrol officers should do, how they should behave and what they should be able to deal with, it was necessary for Peter Bloch, in directing The Urban Institute evaluation of policewomen on patrol in the District of Columbia, to report in the *summary* findings three measures of comparability of assignment, 23 measures of performance, three of citizen attitudes and 13 of police attitudes. This experiment took two years and cost over $300,000.

Few police agencies ever have these levels of funding free of operational commitment. For major evaluations, outside funding is almost always a necessity.

We have set out in simple terms what an agency needs to do to participate effectively in evaluative research. But why should they?

It is common for administrators of all kinds to believe that evaluations of programs they direct are threatening, that such evaluations may cast *them* in bad light if the results are not positive, not just the program. Police chiefs or other police administrators are no exception to this tendency.

Not only that, but there is positive, political potential in starts that have no endings. The value and power of starts must not be underrated. Any study of experienced specialists in bureaucratic survival is likely to show that they understand and make full use of this principle—that starts of new projects, new contracts, almost anything—can be announced with fanfare, can be made to seem important and good simply by rhetoric, and can lead to gains in image, all at relatively little cost since they are often paid for with outside money. Endings can too often be, at best, modest as compared to opening rhetoric, at worst, downright damaging. The thing to do is to start as often as possible, let the project disappear quietly when that money is gone and bury the disappearance even more deeply by new starts. Until recently this has worked well for any administrator who chose, or unwittingly found himself in, this cycle. Now some law enforcement outside funding is tied to evaluation commitments and some of these will be implemented. But the relative power of the start is still a force to reckon with. It does not invite evaluation.

Not only that, but some police administrators who begin well designed, purposeful research in good faith, on matters that they intend to result in real change, responsive to the knowledge they hope to gain, can be disappointed part way through the process. It is natural for operationally oriented people, like police chiefs, to want to move; they live on short time scales, where palpable action counts. Sometimes they get impatient with evaluators who do not know what the results of an experiment show as soon as the last data are collected. It may take as long as a year to analyze and synthesize the vast quantities of data involved in major evaluations. In the meantime, the chief may feel there is a real cost to waiting. It can happen that he has unreal expectations of the knowledge analysts have and the use they can make of it. He may not know that, with rare exceptions, operational judgments about "what happened" in an experiment are still best made by his own operational staff, not by analysts despite the piles of raw data they may have. Their contributions to empirical knowledge come from their ability to analyze and ultimately to understand the meaning of complex data sets. Evaluators, for their part, may feel sympathy for the chief's sense of need and try to give interim indications earlier than they find their knowledge of the facts allows. This situation is a potential source of irritation to both police agency and evaluation staffs. Good feedback loops and patience are needed to avoid or correct unreal expectations of each other by these two very different kinds of people.

Not only that, but we have said that the chief must be prepared when he undertakes to conduct an experiment to discipline people in his own agency if they do not support or if they interfere with maintaining necessary experimental conditions. People have been removed from positions or reassigned. The internal political costs to do that can be high.

Not only that, while no one would expect experimental conditions to be maintained that consciously jeopardized the safety of citizens, and it is understood in every case that a chief will stop an experiment in which the evidence shows that that is taking place, nevertheless the chief is taking risks when he starts an experiment. He risks losing public support of citizens who do not understand what he is doing to assure no significant change in their safety during an experiment. He may feel that he may risk losing support of his city management if results are not favorable. These risks are real. The average tenure of police chiefs in this country is only about three years. Survival is his main preoccupation, and he well knows the whimsical nature of the determinants of his tenure: a replaced mayor or manager or one breaking scandal which catches him with surprise can overbalance precious years of satisfactory performance.

What are the inducements to accept these risks and challenges? Why is it that police agencies have attempted as many as four formal experiments at once? (That, we learned, is three too many even for a department with more management capacity than most. The concentration of attention at the highest level to insure that one major experiment

can be controlled, along with running the department on a day-to-day basis—no small job in itself—dictates attempting only one major experiment at a time.) The forces that lead to doing so must be powerful.

There are environmental ones. The public, the Federal government, community groups, and scholars have been applying pressure for improvements in police civility and effectiveness for about a decade. When a city council tells a chief to show the effectiveness of a practice that has become controversial or abandon it, the chief can become more receptive to formal testing. In that process, elements of his department can see and seize upon the opportunity to plan, secure his approval, get financial support for, and test a different concept of policing, further responsive to the city council's concerns, which changes the role of a patrol officer. In three years, looking back, John Boydstun of System Development Corporation directed evaluations of both the San Diego Field Interrogation and Community Profile experiments and the department is now engaged with us in a most complex and difficult experiment to attempt to measure the relative desirability of one- versus two-officer staffing of patrol cars. The department has committed itself to and is engaged in adopting Community-oriented Policing throughout its patrol force. The chief and the department are looking ahead to planning more tests of patrol practice.

What began largely as a response to environmental pressure is now an accepted mode of working. This has happened in other police agencies too, because there are many in policing, chiefs and others, who feel strongly the need to learn and change and will respond to opportunity. The Foundation sometimes represents such an opportunity. So, internal forces can also be strong.

Foundation funding is another reason. Bringing in external funding can have political value in itself. But, in most instances, Foundation program grants are small compared to the police budgets they might be thought to influence. Foundation funding certainly has facilitated the thoughtful testing of ideas by those police agencies that wish to do so, but, by itself, could not do more than that. Expenditure of $30 million on police research and reform over a period of some eight years cannot be expected to force the changing of an enterprise that will have spent, perhaps, well over $50 billions or more over that time span.

But change, and research and experimentation in policing is going on increasingly. A principal reason seems to be that many leaders in policing have concluded that this is the distinguishing mark of leadership—to be open, to query, to test in a formal sense and then apply what is learned and move forward by such reasoned steps. Others, who wish to be seen as leaders in their own right, are finding that this is the way to do so credibly. They are joining forces with the earlier innovators. This is the basis of the strength that is now showing, despite how much more needs doing.

One caution is due to those who would follow in this excellent path. The definition of success must be fully understood. It is customary for almost any administrator or program manager, including those in policing, once he has decided what to do, to commit himself to the success of the program or practice. He commonly does so in such a way that if it fails, he fails. Hence his uneasiness about evaluation. The leading innovators' approach is different. They focus their attention on the problem to be dealt with and they commit themselves to a *fair test* of the most effective approach they can devise or find at the time. If the test shows that not to be effective, they make changes or apply another technique or practice and test again. They do not fail when a program does not operate or deliver as expected. They only fail if they do not try another approach improved by what they learned in the test.

An evaluator measures the success of an experiment, not in terms of whether the outcomes were as expected or hoped for by the agency, but rather, in terms of whether he knows what happened. (This difference can lead to friction.) The only failure an experiment can have is not to know. The leading innovators in policing have adopted some of that philosophy. Innovators in other kinds of enterprise may find it useful.

Biographical Sketches

Jan Acton is an economist with Rand Corporation, currently working on energy problems. He did his undergraduate work at San Diego State College, and completed his Ph.D. at Harvard in economics. His doctoral thesis was an assessment of strategies for treating victims of heart attack. He also analyzed several measures for valuing the lives that might be saved by emergency interventions.

William Bieck is Principal Investigator on the Response Time Analysis Study, a five-year project funded through the National Institute of Law Enforcement and Criminal Justice, for the Kansas City, Missouri, Police Department. Prior to joining the Kansas City, Missouri, Police Department, he was employed by the Police Foundation as an Observer on the Kansas City Preventive Patrol Study. This experience enabled him to monitor patrol operations first hand, having accompanied over 50 officers across all watches for a period of 14 months. Before his employment with the Police Foundation, he was an Assistant Professor of Sociology at Nebraska Wesleyan University in Lincoln, Nebraska, and an Instructor in the Department of Law Enforcement and Correction for the University of Nebraska at Omaha; a position he held for seven years. Mr. Bieck has a B.S. in Psychology and a M.A. in Sociology.

Professor Robert F. Boruch is Director of the Methodology and Evaluation Research Division, Psychology Department, Northwestern University, and current President of the Council for Applied Social Research. He is a coauthor of *Social Experimentation* and an editor of *Experimental Tests of Public Policy;* he has published over thirty journal articles dealing with methodological, managerial, and ethical problems in research. Dr. Boruch is a member of advisory panels of the National Academy of Sciences, the American Psychological Association, and consults frequently for Federal agencies on research planning and design.

Russell D. Clark III is a social psychologist and Associate Professor at Florida State University. He did his undergraduate work at Tarkio College and his graduate work at the University of Kansas. Aside from work on attitude measurement he has studied the influence of groups on decision making and the factors influencing helping behavior.

Linda Victor Esrov's educational background is in experimental psychology. She received her B.A. from Temple University and a Ph.D. from Northwestern University. She also completed a two-year post-doctoral fellowship in evaluation research with Lee Sechrest at Florida State University and has been involved in a number of evaluation projects concerning emergency medical services.

Lieut. Colonel Lester Harris had been a member of the Kansas City, Missouri Police Department for twenty-two (22) years and is currently assigned as Assistant Chief of Police. Past assignments include patrol, instructor of the Police Academy, Commander of Planning and Research, Commander of a patrol division, Commander of an investigations division, Assistant Commander of both the Administration Bureau and the Operations bureau and Commander of the Services Bureau. He is a 1968 graduate of the Southern Police Institute and has attended Central Missouri State University, majoring in Criminal Justice Administration.

George Kelling received his BA degree from St. Olaf College, his Master's degree in social work from the University of Wisconsin-Milwaukee, and his Ph.D. in social work from the University of Wisconsin-Madison. Prior to beginning his work with the Police Foundation, where he is currently employed, he was involved in probation and parole activities, and institutional work with aggressive youngsters. He was also an Assistant Professor of Social Work at the University of Wisconsin-Milwaukee. Since joining the Police Foundation in 1971, Kelling has worked on evaluation studies in Dallas & Kansas City, and is now involved in a large scale study of police foot patrol in several cities in New Jersey.

Joseph Lewis is Director of Evaluation at the Police Foundation in Washington, D.C. His first degree was from the University of Maine in Electrical Engineering, but he received a subsequent Master's degree in Economics and Business Administration. After working as an engineer for Consolidated Edison and the U.S. Navy, Lewis had a brief and successful career in private industry before joining the Office of the Secretary of Defense. From there he went to the Institute for Defense Analysis where he developed and directed Command and Control Activities of the Weapons System Evaluation Group. In 1968 he joined the Urban Institute staff as Director of the Urban Governance Research Program and remained there until 1971 when he assumed his present position at the Police Foundation.

Lee Sechrest received all three of his academic degrees from the Ohio State University, with a major in clinical psychology. He taught for two years at Pennsylvania State University before going to Northwestern University, where he remained for fifteen years. During his tenure at Northwestern, Sechrest became interested in program evaluation and in health services research and was instrumental in developing training programs in both those areas. He moved to Florida State University in 1973, where he is Professor of Psychology. He is a past member of the Health Services Research Study Section and is currently involved in work on assessing performance of emergency medical technicians.

Current NCHSR Publications

The following National Center for Health Services Research publications are of interest to the health community. Copies are available on request to NCHSR, Office of Scientific and Technical Information, 3700 East-West Highway, Room 7-44, Hyattsville, Maryland 20782 (tel.: 301/436-8970). Mail requests will be facilitated by enclosure of a self-adhesive mailing label.

PB and HRP numbers in parentheses are order numbers for the National Technical Information Service (NTIS), Springfield, Virginia 22161 (tel.: 703/557-4650). Those publications which are out of stock are indicated as available from NTIS. Prices may be obtained from the NTIS order desk on request.

Research Digests

The *Research Digest Series* provides overviews of significant research supported by NCHSR. The series describes either ongoing or completed projects directed toward high priority health services problems. Issues are prepared by the principal investigators performing the research, in collaboration with NCHSR staff. Digests are intended for an interdisciplinary audience of health services planners, administrators, legislators, and others who make decisions on research applications.

(HRA) 76-3144 Evaluation of a Medical Information System in a Community Hospital (PB 264 353)

(HRA) 76-3145 Computer-Stored Ambulatory Record (COSTAR) (PB 268 342)

(HRA) 77-3160 Program Analysis of Physician Extender Algorithm Projects (PB 264 610)

(HRA) 77-3161 Changes in the Costs of Treatment of Selected Illnesses, 1951-1964-1971 (HRP 0014598)

(HRA) 77-3163 Impact of State Certificate-of-Need Laws on Health Care Costs and Utilization (PB 264 352)

(HRA) 77-3164 An Evaluation of Physician Assistants in Diagnostic Radiology (PB 266 507)

(HRA) 77-3166 Foreign Medical Graduates: A Comparative Study of State Licensure Policies (PB 265 233)

(HRA) 77-3171 Analysis of Physician Price and Output Decisions

(HRA) 77-3173 Nurse Practitioner and Physician Assistant Training and Deployment

(HRA) 77-3177 Automation of the Problem-Oriented Medical Record

Research Summaries

The *Research Summary Series* provides rapid access to significant results of NCHSR-supported research projects. The series presents executive summaries prepared by the investigators at the completion of the project. Specific findings are highlighted in a more concise form than in the final report. The *Research Summary Series* is intended for health services administrators, planners, and other research users who require recent findings relevent to immediate problems in health services.

(HRA) 77-3162 Recent Studies in Health Services Research, Vol. I (July 1974 through December 1976) (PB 226 460)

(HRA) 77-3176 Quality of Medical Care Assessment Using Outcome Measures

Policy Research

The *Policy Research Series* describes findings from the research program that have major significance for policy issues of the moment. These papers are prepared by members of the staff of NCHSR or by independent investigators. The series is intended specifically to inform those in the public and private sectors who must consider, design, and implement policies affecting the delivery of health services.

(HRA) 77-3182 Controlling the Cost of Health Care (PB 266 885)

Research Reports

The *Research Report Series* provides significant research reports in their entirety upon the completion of the project. Research Reports are developed by the principal investigators who conducted the research, and are directed to selected users of health services research as part of a continuing NCHSR effort to expedite the dissemination of new knowledge resulting from its project support.

(HRA) 76-3143 Computer-Based Patient Monitoring System (PB 266 508)

(HRA) 77-3152 How Lawyers Handle Medical Malpractice Cases (HRP 0014313)

(HRA) 77-3159 An Analysis of the Southern California Aribtration Project, January 1966 through June 1975 (HRP 0012466)

(HRA) 77-3165 Statutory Provisions for binding Arbitration of Medical Malpractice Cases (PB 264 409)

(HRA) 77-3184 1960 and 1970 Spanish Heritage Population of the Southwest by County

(HRA) 77-3188 Demonstration and Evaluation of a Total Hospital Information System

(HRA) 77-3189 Drug Coverage under National Health Insurance: The Policy Options

(HRA) 77-3191 Diffusion of Technological Innovation in Hospitals: A Case Study of Nuclear Medicine (in preparation)

Research Management

The *Research Management Series* describes programmatic rather than technical aspects of the NCHSR research effort. Information is presented on the NCHSR goals, research objectives, and priorities; in addition, this series contains lists of grants and contracts, and administrative information on funding. Publications in this series are intended to bring basic information on NCHSR and its programs to research planners, administrators, and others who are involved with the allocation of research resources.

(HRA) 76-3136 The Program in Health Services Research (Revised 9/76)

(HRA) 77-3158 Summary of Grants and Contracts, Active June 30, 1976

(HRA) 77-3167 Emergency Medical Services Systems Research Projects (Active as of June 30, 1976) (PB 264 407, available NTIS only)

(HRA) 77-3179 Research on the Priority Issues of the National Center for Health Services Research, Grants and Contracts Active on June 30, 1976

(HRA) 77-3183 Recent Studies in Health Services Research, Vol. II (CY 1976)

Research Proceedings

The *Research Proceedings Series* extends the availability of new research announced at key conferences, symposia and seminars sponsored or supported by NCHSR. In addition to papers presented, publications in this series include discussions and responses whenever possible. The series is intended to help meet the information needs of health services providers and others who require direct access to concepts and ideas evolving from the exchange of research results.

(HRA) 77-3138 Women and Their Health: Research Implications for a New Era (PB 264 359, available NTIS only)

(HRA) 77-3150 Intermountain Medical Malpractice (PB 268 344, available NTIS only)

(HRA) 77-3154 Advances in Health Survey Research Methods

(HRA) 77-3181 NCHSR Research Conference Report on Consumer Self-Care in Health

(HRA) 77-3186 International Conference on Drug and Pharmaceutical Services Reimbursement

BIBLIOGRAPHIC DATA SHEET	1. Report No. NCHSR 78-46	2.	3. Recipient's Accession No.
4. Title and Subtitle EMERGENCY MEDICAL SERVICES: RESEARCH METHODOLOGY; CONFERENCE HELD IN ATLANTA, GEORGIA, SEPTEMBER 8-10, 1976; NCHSR Research Proceedings Series			5. Report Date December 1977 6.
7. Author(s) Lee Sechrest (conference director/editor)			8. Performing Organization Rept. No.
9. Performing Organization Name and Address Jacksonville Experimental Health Delivery System, Inc. 1045 Riverside Avenue (Suite 275) Jacksonville, Florida 32204			10. Project/Task/Work Unit No. --- 11. Contract/Grant No. HSM 110-72-314
12. Sponsoring Organization Name and Address DHEW, PHS, OASH, National Center for Health Services Research 3700 East-West Highway, Room 7-44 (STI) Hyattsville, Maryland 20782 (Tel.: 301/436-8970)			13. Type of Report & Period Covered Proceedings Sept. 8-10, 1976 14.

15. Supplementary Notes

DHEW Pub. No. (PHS) 78-3195.

16. Abstracts

The focus of this conference was the importance of systematic research in evaluating the Emergency Medical Services system and administrative functions. Presentations made at the conference and compiled in this document deal with a range of conceptual and methodologic issues. Particular attention is given to the opposing yet mutually dependent roles of the administrator/evaluator. Several papers presenting aspects of research conducted in a police setting offer an instructive analogy to emergency medical services systems.

~~17. Key Words and Document Analysis. 17a. Descriptors~~

15. Supplementary Notes (Continued)

NCHSR publication of research findings does not necessarily represent approval or official endorsement of research findings by the National Center for Health Services Research or the Department of Health, Education, and Welfare.

17b. Identifiers/Open-Ended Terms

Health services research
Emergency medical services
Research methodology

17c. COSATI Field Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Releasable to the public. Available from National Technical Information Service, Springfield, VA 22161 (Tel.: 703/557-4650)	UNCLASSIFIED	Est. 125
	20. Security Class (This Page) UNCLASSIFIED	22. Price

FORM NTIS-35 (REV. 10-73) ENDORSED BY ANSI AND UNESCO. THIS FORM MAY BE REPRODUCED USCOMM-DC 8265-P74